26942

National Library of Canada | Bibliothèque nationale du Canada

CANADIAN THESES ON MICROFICHE | THÈSES CANADIENNES SUR MICROFICHE

NAME OF AUTHOR/*NOM DE L'AUTEUR* HESSEL H. VANDERDUIM

TITLE OF THESIS/*TITRE DE LA THÈSE* THE WALKER PROBLEM IDENTIFICATION CHECKLIST: A COMPARISON BETWEEN REGULAR CLASS AND JUNIOR ADAPTATION CLASS STUDENTS

UNIVERSITY/*UNIVERSITÉ* UNIVERSITY OF ALBERTA

DEGREE FOR WHICH THESIS WAS PRESENTED/ *GRADE POUR LEQUEL CETTE THÈSE FUT PRÉSENTÉE* MASTER OF EDUCATION

YEAR THIS DEGREE CONFERRED/*ANNÉE D'OBTENTION DE CE GRADE* 1975

NAME OF SUPERVISOR/*NOM DU DIRECTEUR DE THÈSE* GERARD M. KYSELA.

Permission is hereby granted to the NATIONAL LIBRARY OF CANADA to microfilm this thesis and to lend or sell copies of the film.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

L'autorisation est, par la présente, accordée à la BIBLIOTHÈQUE NATIONALE DU CANADA de microfilmer cette thèse et de prêter ou de vendre des exemplaires du film.

L'auteur se réserve les autres droits de publication; ni la thèse ni de longs extraits de celle-ci ne doivent être imprimés ou autrement reproduits sans l'autorisation écrite de l'auteur.

DATED/*DATÉ* August 12/75. SIGNED/*SIGNÉ* Hessel H Vanderduim

PERMANENT ADDRESS/*RÉSIDENCE FIXE* 967 7½ Avenue East
Owen Sound, Ontario
N4K 2Y9.

NL-91 (3-74)

THE UNIVERSITY OF ALBERTA

THE WALKER PROBLEM BEHAVIOR IDENTIFICATION CHECKLIST: A COMPARISON BETWEEN REGULAR CLASS AND JUNIOR ADAPTATION CLASS STUDENTS

by

HESSEL H. VANDERDUIM

©

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF EDUCATION

DEPARTMENT: EDUCATIONAL PSYCHOLOGY

EDMONTON, ALBERTA

FALL, 1975

PREVIOUSLY COPYRIGHTED MATERIAL IN
APPENDICES I AND II (LEAVES 101-112)
NOT MICROFILMED.

PLEASE CONTACT THE UNIVERSITY FOR FURTHER
INFORMATION:

BOARD OF RESEARCH AND GRADUATE STUDIES
UNIVERSITY OF ALBERTA
EDMONTON, ALBERTA

THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research, for acceptance, a thesis entitled THE WALKER PROBLEM BEHAVIOR IDENTIFICATION CHECKLIST: A COMPARISON BETWEEN REGULAR CLASS AND JUNIOR ADAPTATION CLASS STUDENTS submitted by HESSEL H. VANDERDUIM in partial fulfilment of the requirements for the degree of Master of Education.

Gerard W. Kysela
Supervisor

Date August 11, 1975

DEDICATION

To Barbara, without whose love
and encouragement this would
not have been possible

ABSTRACT

Students from regular classes and Junior Adaptation classes were rated by their teachers on the Walker Problem Behavior Identification Chacklist (WPBIC). A total of 188 subjects, divided evenly between the two groups, were used and the results were compared.

On the full scale and on 4 of the 5 subscales (Acting-out, Distractability, Disturbed Peer Relations and Immaturity) the mean scores of the two groups showed differences significant beyond the .001 level of confidence. The remaining scale (Withdrawal) showed differences between the groups, significant beyond the .01 level of confidence but the results here were felt to be seriously compromised by a nested teacher-related effect on scores. A similar nested teacher related effect was found in relation to the full scale but was overwhelmed by real differences between the groups.

Frequency polygons for the full scale and all subscales were generated and the amount of overlap around Walker's cut-off points for problem behavior was rated. On all but Scale II (Withdrawal) substantially more subjects in the Junior Adaptation classes were rated as possessing problem behavior than subjects in the regular class group. It was noted that the hypotheses in regard to these polygons

may have been too restrictive in their setting of limits.

Groups matched for age and sex were selected from the two source groups and compared for differences of scores. On all but Scale II (Withdrawal) and Scale V (Immaturity) differences beyond the .001 level of confidence were noted. On Scale V (Immaturity) differences beyond the .05 level of confidence were noted while on Scale II (Withdrawal) no significant differences were found. It was determined that these groups were similar to the source groups.

The Junior Adaptation sub-group used above was compared to a sub-group from Walker's original sample, this sub-group being identified as possessing disturbed behavior. The sub-group from the present study and Walker's subgroup appeared to be very similar in that there was no statistically significant difference in their mean scores or variance.

Small, negative correlations were found between subjects' ages and checklist scores and between length in Junior Adaptation classes and checklist scores. These results were not considered significant, however.

The results of the findings were discussed and certain conclusions were drawn regarding the usefulness of the instrument as a device for screening students for evidence of problem behavior.

ACKNOWLEDGEMENTS

I Have a lot of people to thank for the completion of this thesis. Many must go unnamed but I have appreciated their support and encouragement.

I particularly want to thank:

Dr. G. Kysela for deeping me on the straight and narrow during the preparation of this thesis and Dr. L. Stewin for his help and encouraging words.

Prof. S. Campbell for serving on my committee.

Dr. Charrette, the principals, teachers and students of the Edmonton Public School Board for giving so freely of their time so that this research could be done.

Ron Aubin, without whose patience and way with computers I'd still be lost in a sea of data.

Mrs. Peggy Mulligan, who spent what must have been weeks putting this manuscript into recognizable form.

Barbara, my wife, whose courage and selflessness gave me the chance to give up a steady job for the uncertainty of university life and the opportunity to further my education.

TABLE OF CONTENTS

Title	Page
I. INTRODUCTION	1
Background and Problem	1
Purpose of the Study	4
II. REVIEW OF LITERATURE	6
A Model for the Development of Behavior	6
Review of Relevant Literature	13
Ratings and the Criteria of Ratings	23
Factors Relating to Rates	24
III. RATIONALE	26
Definition of Terms	27
Hypotheses	30
Design	31
IV. METHODS	34
Subjects	34
The Instrument	35
Procedures	39
(a) Raters	40
(b) Data Collection	40
Data Analysis Techniques	41
V. RESULTS	45

Title Page

VI. DISCUSSION AND CONCLUSIONS 81

Discussion 81

Conclusions and Implications 92

REFERENCES 96

APPENDICES . . . . . .

1. Walker Problem Behavior Identification Checklist Manual 101

2. Sample - Walker Problem Behavior Identification Checklist 111

LIST OF TABLES

Table		Page
1.	Hierarchal Analysis - Full Scale	46
2.	Hierarchal Analysis - Scale I	48
3.	Hierarchal Analysis - Scale II	49
4.	Hierarchal Analysis - Scale III	51
5.	Hierarchal Analysis - Scale IV	52
6.	Hierarchal Analysis - Scale V	54
7.	Comparisons Means - Full and Sub-Scales	56
8.	Comparisons of Variance - Full Scale Only	57
9.	Comparisons of Means - Matched Groups	70
10.	Comparison of Means - Group II to Walker (1970)	74
11.	Comparison of Variance - Group II to Walker (1970)	76
12.	Comparison of Means - Matched Groups to Source Groups	77

LIST OF FIGURES

Figure		Page
1.	Group Structure for a School	33
2.	Distribution of Full Scale Scores	61
3.	Distribution of Scale I Scores	63
4.	Distribution of Scale II Scores	64
5.	Distribution of Scale III Scores	65
6.	Distribution of Scale IV Scores	66
7.	Distribution of Scale V Scores	68

I INTRODUCTION

Background and Problem

When does a child's negative behavior cease to be merely annoying and begin to be problematic? Within this seemingly innocuous question are lodged certain definitional problems that affect the attempt to find an answer.

Of immediate concern is the word "negative." In Allee (1958) the form "negative" is defined as ". . . expressing denial, prohibition, or refusal; lacking positive qualities; not positive" (p. 250). However, behaviors that may annoy a teacher in one classroom eg. attention seeking, extreme perfectionism, even aggressiveness, may not be perceived as annoying by another teacher or the child's parents. It may be that, at the time, the annoyed teacher is hyper-critical, impatient, or simply busy and perceives the behavior as annoying. It may also be that the behavior in question is, in fact, serious enough to warrant outside attention, but not recognized as such by the parents' untrained eyes or the teacher's differing standards.

Contained within the question is also the matter of defining the word "problematic." What constitutes problem behavior? Is it merely the manifestation of overt "negative" behaviors? Certainly, the child whose behavior is overtly aggressive, or immature, or reflective of distractability,

may be considered to have a behavior problem. However, should we not, also, consider as problematic, the behavior of the child who has sat in the classroom from September to May without uttering a sound. Perhaps, because this child has never defied the teacher, or struck another student, the teacher assumes that there is no problem because the child makes no demands on the teacher's time.

The point to be made here is that the evaluation and identification of problem behavior can be a very subjective and difficult matter. The concerned and dedicated teacher may be reluctant to "label" a child on the basis of a subjective evaluation, for fear that the "label" may follow the child even if it no longer applies. It is not hard to visualize the potentially devastating long term effects that mis-identification or faulty "labelling" of a child could have. On the other hand, a child whose problem behavior goes unchecked because of a teacher's fear of mis-identification, will not only suffer personally but cause others to suffer as well.

The need, then, exists for a relatively simple, comprehensive, valid and reasonably reliable instrument which can help to determine if a particular child's pattern of behavior warrants action by outside personnel, or whether this behavior can be dealt with successfully by the regular class teacher. Also, if factors related to the particular behavior type can be isolated, then decisions can be made

as to a strategy for remediation. Such an instrument would, of course, form only a part of the process for selecting children with serious behavioral problems who would require specialized instruction and remediation.

The Walker Problem Behavior Identification Checklist, developed in 1970, has been claimed to be useful in identifying children with behavioral problems. Based on a norming sample of 534 Grade 4, 5 and 6 students, the instrument is said to identify behavior problems along 5 dimensions (acting-out, withdrawal, distractability, disturbed peer relations, immaturity). Whether or not the instrument is generalizable to the North American population as a whole remains to be adequately proven.

If, in the present study, Walker's checklist were to distinguish between children in regular classes and those in classes for children with behavior disorders, its usefulness as a device for detecting behavioral problems could be considered. As such, the instrument could then be used by teachers to assist in making decisions regarding their students' behavior.

If, on the other hand, the instrument were to fail to distinguish between children in the two groups, serious questions could be raised concerning the instrument's usefulness. In fact, the instrument should perhaps then be subject to a thorough re-examination in terms of its standardization, validity, and design.

Alternative to weakness in the instrument, in the event of failure to discriminate, would be a consideration of weakness in the selection process used to place the behaviorally disturbed children. On the surface, a weakness in this process would seem to be a relatively remote possibility, since these children have been subjected to extensive screening before placement, and each case has been carefully considered. This would not, however, rule out the possibility of a problem in this area.

The process of successfully identifying children with behavior problems, and determining the type and severity of the problem, has been one of some concern. An instrument which may be of some use in this process, has been available for some time, but could benefit from investigation as to its exact usefulness and limitations.

The main focus of this study will be to investigate the instrument's ability to distinguish between children who have been identified as possessing some behavioral disorders and those not so identified.

Purpose of the Study

The present study will attempt to determine the extent to which the Walker Problem Behavior Identification Checklist can differentiate between children in a Grade 4, 5 or 6 setting who have not been identified, formally, as possessing behavior problems and children in Junior

Adaptation classrooms, who are approximately the same age, but have been identified as possessing problem behavior. Since the checklist claims the ability to discriminate between problem and non-problem behavior, it should show differences of a significant nature between these two groups of children.

In addition to determining whether or not this instrument can detect overall differences between the groups, it is of interest to determine the percentage of children in each group that, according to the Checklist, are misclassified. That is, what percentage of children in the regular classes are identified by the criteria of the checklist as possessing problem behavior and, of possibly greater importance, what percentage of children in Adaptation classes are identified as not possessing problem behavior. Obviously, an inordinate percentage of misclassified children would raise questions about the discriminatory properties of the checklist, or perhaps more seriously, the procedures used to select students for inclusion in Adaptation classes in the local system.

The ability or inability to differentiate between the two groups should yield an estimate of construct validity, in the form of contrasted groups validity.

II REVIEW OF LITERATURE

A Model for the Development of Behavior

Before an examination of problem behavior can be initiated, a model for behavior should first be established. For the purpose of this study, behavior will be considered to develop in terms of the Social Learning model outlined by Bandura and Walters (1963) and modified somewhat by Bandura (1973).

The essence of Social Learning theory centres around the notion that behavior is a learned, rather than innate phenomenon. While not excluding the possibility of spontaneous behavior, based on available cues in the immediate environment, Bandura and Walters suggest that much, if not most, learning occurs from observing a model perform or hearing a model verbalize the performance of a behavior. Especially if the model is rewarded for the particular behavior, the learner tends to learn much more rapidly and easily than if no model was present. According to this view, the model's behavior is copied by a series of successive approximations until the behavior is mastered.

Bandura (1973) qualifies this position somewhat to account for behavior not being learned even under all possible favorable conditions. He states:

> . . . Exposure to models, even prestigious ones, does not automatically produce matching performances. In any given instance absence of imitative behavior may result from faulty observation, retention losses due to inadequate symbolic representation and rehearsal, motor deficiencies, or simply unwillingness to perform the exemplified behavior because of its unfavorable consequences (p. 72).

This somewhat enlarged view accounts more adequately for the absence of learning of modelled behavior than did the earlier position. It grants more scope for individual differences in the learning process and attempts to specify some possible reasons behind the absence of learning in some cases.

Bandura (1973) goes on to point out certain effects that modelled behavior can have on behavior in the observer:

> First . . . observers can acquire new patterns of behavior through observation. A second major function of modelling influences is to strengthen or weaken inhibitions of behavior that observers have previously learned. Inhibitory and disinhibitory effects are largely determined by observation of rewarding and punishing consequences accompanying model's responses. The actions of others also serve as social prompts that facilitate similar behavior in observers. Response facilitation effects can be distinguished from observational learning and disinhibition by the fact that no new responses are acquired, and the appearance of analogous actions is not attributable to weakening of inhibitions because the behavior is socially acceptable and, hence, unencumbered by restraints (pp. 68-69).

The model can, thus, show new behavior, strengthen or weaken inhibitions toward behavior by the observer, or present social cues or situations where a particular behavior is acceptable. The role of the model in Social Learning theory can be considered to be crucial to the development

of behavior.

Once the behavior has been learned or modelled its survival depends upon the effect of reinforcement or reward. A reinforced-rewarded-behavior can be expected to recur if a situation like the one in which the behavior was learned recurs. A behavior that is reinforced on each occurence or emission can be learned very quickly, but can also be extinguished (unlearned, for want of a better word) if the reinforcer is no longer presented. Behavior whose reinforcement occurs irregularly in terms of time between presentation of the reinforcer, or in terms of the number of emissions of the behavior, or a combination of the two above, will be learned more slowly that behavior reinforced each time it occurs. However, behavior learned in this way also resists extinction to a high degree and can be expected to persevere over a considerable length of time, even in the absence of any reinforcement.

This basic position has one major qualification expressed by Bandura (1973):

> Response consequences (i.e. reinforcement contingencies) . . . have weak effects on behavior when the relationship between one's actions and outcomes is not recognized. On the other hand, awareness of conditions of reinforcement typically results in rapid changes in behavior, which is indicative of insightful functioning. People who are aware of what is wanted and value the contingent rewards, change their behavior in the reinforced direction; those who are equally aware of the reinforcement contingencies but who devalue the required behavior or reinforcers show little change; those who remain unaware achieve, at best, small increment in performances even though the appropriate responses are reinforced whenever they occur (p. 50).

Two salient points are raised here that distinguish Social Learning theory from the strict behaviorist view one might expect from Skinner (see Baldwin, 1967). First, as compared to the behavioral view, cognition is recognized as a key aspect of the learning process. That is, the learner must be aware of the reward contingencies affecting a learning situation if he is to make significant progress. Second, valuation becomes a factor. Even if the learner is aware of the reward contingencies, he must value the rewards before he will make gains. This valuing of the reward may be termed motivation, for want of a better word, and differs from the behavioral concept of motivation which sees it as a state of need or deprivation. The Social Learning view adds a level of sophistication to the concept, again through its cognitive dimension, that seems to be lacking in the other view.

Social Learning theory views the acquisition and emission of learned behavior in terms of generalization and discrimination to relate how one behavior may or may not be used in a variety of differing though similar situations. To generalize a behavior one emits a particular behavior in a situation whose cues are very similar to that in which the behavior was first learned. As these cues or signals for behavior are more dissimilar, the particular behavior is less likely to be emitted. If this process breaks down, an individual may overgeneralize and emit a behavior totally inappropriate to the situation. By the same token, and

concurrent to a generalization process, a very strict process of discrimination must occur so that individual can determine whether or not a particular behavior is appropriate to inappropriate to a given situation. If this process does not function properly or if the cues or modelled behavior are inappropriate or misunderstood, inappropriate behavior will likely result. In a very real sense, generalization and discrimination are two sides of the same coin. A breakdown in either process will affect the other. In fact, Bandura (1973) tends to refer to the two as parts of the same whole.

Various other effects operate to influence the learning of the individual. An individual who is highly dependent, that is, one who constantly looks to outside sources for reinforcement will be much more susceptible to social influence than a person who is highly independent in his actions. It should be noted here that both dependency and independence are, in themselves, learned behaviors and as such can be altered through behavior modification as can most behavior according to Bandura and Walters.

Sex differences form an integral part of the social learning process. Social demands have traditionally been different for girls than for boys and the extent to which these demands are learned will affect the individual's response to a given set of situational cues.

The sex of the individual who is the model for

behavior can also be of considerable importance. The modelling effect, i.e. extent of imitation, will be greater if the model is the same sex as the learner.

The performance of learned behavior is also more likely if the model is a "high prestige" individual. If the model is held in high regard by the learner, then his behavior is more likely to be imitated. Again the concepts of cognition and valuation are considered here as being quite important.

If there is a state of emotional arousal, i.e. some form of excitation, anxiety, or interest, learning seems to be facilitated over a state of non-arousal or neutral affect to the situation. However, this aroused state is only a facilitator of learning within certain limits. If an optimal level of arousal is passed, then the aroused state can, conceivably, block learning because of the individual's concentration on the source of arousal and its elimination, rather than the learning situation.

All of the above mentioned factors are, according to Bandura and Walters, based on some prior experience or previous learning in a similar setting. This may also be said of conflict, i.e. approach avoidance situations where two mutually exclusive options are present and of displacement, i.e. the transfer of a desired response from one object to another.

Social Learning theory has undergone some evolutionary

change in regard to views on punishment and non-reward. Whereas Bandura and Walters (1963) saw punishment as inhibiting behavior without removing it from the behavioral repertoire and non-reward as extinguishing behavior, Bandura (1973) presents a somewhat more thoroughly considered view:

> There are two principal ways in which negative sanctions inhibit forbidden actions. Repeated punishment for aggressing toward certain persons places or things endows them with fear arousing value. As a result, inclinations to aggress toward these threats evokes fear, which motivates inhibitory controls.
>
> The effectiveness of punishment in controlling behavior is determined by a number of factors. Of special importance is the level of reward achieved through (aggressive) conduct and the availability of alternative means of securing the desired goals. The likelihood that aggression will be punished, the nature, severity and duration of the aversive consequences, and the time elapsing between aggressive actions and negative outcomes also determine the suppressive power of punishment. Additionally, the level of instigation to aggression and the characteristics of the prohibitive agents influence how aggressors will respond to being punished (pp. 221-222).

While the above statements specify aggressive behavior, various forms of undesirable behavior could be substituted where the words "aggression" or "aggressive" appear and still apply.

The appeal of Social Learning theory lies in the fact that its emphasis on imitative learning of a model's behavior and the importance of schedules of reinforcement (the manner and timing of reinforcers) applies equally well to both normal and problem behavior.

It is within the Social Learning framework that

Walker's Problem Behavior Identification Checklist appears to have its foundation. The model for behavior acquisition and elimination should give some indication of the developmental nature of behavior acquisition and change. As noted above the model is derived primarily from Bandura and Walters (1963, pp. 1-32), except where otherwise indicated, and has been incorporated primarily to indicate the author's general position on behavior theory.

This study is not overly concerned with the acquisition of behavior, but rather with determining whether or not certain operational descriptions of behavior are useful in describing problem behavior.

Review of Relevant Literature

A number of sources in the current literature on teacher rating of student behavior and attributes will be examined briefly. An attempt has been made to utilize recent articles as much as possible. Since the matter of teacher rating of pupil behavior is central to this study, this will form the major focus of this review.

In an apparent attempt to streamline the process of identifying emotionally disturbed children at the elementary level, Maes (1966) undertook a study which showed that emotionally disturbed children could be identified as effectively through the use of a teacher rating scale and a group I.Q. test as by a battery of measures including

mathematics achievement, reading achievement, a modified sociometric technique (namely class play) and a self concept inventory. The evidence, here, suggests that if a teacher has access to an objective rating instrument, and is trained in the use of it, his classroom observations can be effectively used as a major means of identifying behavioral attributes. The results of this study, though encouraging and supportive of Maes' hypothesis, do not appear to have been validated by further research either by Maes' or any other researcher. Without the support of successful replication, Maes' research, though interesting, is of limited usefulness.

Ebbeson (1968) studied kindergarten teachers' rankings of their students' later academic achievement. It was found that the teachers predicted quite accurately the academic achievement of these students in the early grades of school. The same result was claimed in the prediction of future achievement with two successive kindergarten classes. The successive repetition of Ebbeson's initial results lends some support to his conclusion about the effectiveness of teacher prediction in this case. However, kindergarten teachers do not live in a vacuum and though it was not discussed by Ebbeson, it is entirely possible that these teachers could have passed their views on students to successive teachers, either orally or by written comment, thus biasing the expectations for academic achievement on

the part of those later teachers. This view is entirely speculative, but, given the nature of teacher to teacher communication and the nature of cumulative student records, is one that should be considered.

In an observational study of 10 normal children and a larger group of children with behavior disorders, Werry and Quay (1969) presented evidence to suggest that a method of direct behavioral observation in the classroom is reliable, that it discriminates between normal and disturbed behavior and gives information on the nature of the maladjustment. This work, based on largely individual items of observed behavior suggests that real differences are observable between normal and behaviorally disturbed children. An obvious problem here is the relatively small size of the group identified as normal. A group of 10 subjects can do little more than to suggest general trends and can hardly be used effectively as a standard for normalcy.

Bryan and Wheeler (1972) found that systematic observation of learning disabled children revealed that these children spent significantly less time in task oriented behavior than did non learning disabled children. They stressed, however, the importance of knowing what to look for. Even though a child was looking at a book, something most teachers would consider on-task behavior, he might well not have been reading it. The looking without reading would be off-task behavior and Bryan and Wheeler were

careful to point out that careful and directed observation techniques were needed to successfully perceive behaviors accurately. This seems to be a crucial point and may provide a key to the sometimes inconsistent research results obtained in studying observations by one individual on others. Not only must the observer be sensitive to the group or individual under observation, but, to be effective, he must also know what he is looking at and for. In this area, a structured observational guide would have its primary value.

In a separate study with McGrady, Bryan (1972) analyzed teacher ratings of 183 boys labelled as having learning problems and 176 normal learners. The analysis indicated that teachers consistently rated problem learners lower on each area of the scale used than they did normal learners. Validity was established by comparing the groups identified by the Pupil Behavior Rating Scale with reading and WISC vocabulary scores. On each measure, the learning disabled group scored significantly lower than the normal children. The conclusion formed from this study was that a teacher checklist, in this case the PBRS, could provide an efficient and economical measure for use in screening for learning disability. The authors did, as a cautionary note, suggest further study of the validity of the PBRS and of the basis upon which teachers make their discriminations. This suggestion for further research does not necessarily detract from the value of what appears as a well planned and

executed study which concludes in favor of the utility of a teacher rating scale.

Bullock and Brown (1972) compared reacher reported behavior disorders of 1189 special education students to results on the Behavior Dimensions Rating Scale. A high correlation was found between factors on the BDRS and problems stated as serious by the 112 teachers involved in the research. The findings were used to conclude that teachers appear to have the ability to observe and judge student behavior patterns effectively. The sample size used here seems to lend an air of authority to the study, although the matter of exacting validation of the BDRS, or the lack of it, remains a problem. This problem pervades much of the research involving the use of checklists.

Cowgill, Friedland and Shapiro (1973), in a study using 37 kindergarten boys who had been identified as being learning disabled by the Massachusetts State Department of Education and 37 "normal" kindergarten boys, found that their teachers' evaluations differed significantly on all but one of 7 trait categories and on all general behavior categories used in the study. The results were taken as evidence for the value of teacher reports in identifying learning disabled children. No mention was made as to whether or not the teachers were aware of how each child was "labelled." If they were unaware of a child's classification this research could be considered as useful

support for teacher awareness of pupil attributes. If however they were aware of the children's classification, this knowledge could very easily have biased the observations made. The bias, if present, could tend to lead the teacher to perceive behaviors in such a way as to support the classification. This doubt somewhat compromises the significance of this study.

Garner and Bing (1973), in a study examining differences in pupil-teacher contacts, attempted to correlate verbal teacher-pupil exchanges and teacher ratings of pupils. Students between 7 and 8 years old, from 7 classes, were used. The finding of interest to this study was the high degree of agreement in teacher' ratings of specific pupil attributes, regardless of the amount of contact. It is left to speculation as to whether this agreement reflects similarities in overall attributes of students, or simply similarities in the perceptions of a group of teachers.

Hammet and Batchelor (1973) described the advantages of a behavioral rating questionnaire which parents and teachers could complete to provide more precise and comprehensive data than that obtainable by routine clinical observation. Again the point was made (as by Bryan and Wheeler (1972) and Bryan and McGrady (1972)) that the questionnaire provided direction and structure on which to base observations. In this way, they claimed, the precision available through intensive interaction with the subject

could be maximized and subjective judgments minimized. This interaction effect is certainly valuable in the same way that the structure of the questionnaire is valuable, if that questionnaire has been adequately validated. In this case, the validation of the questionnaire does not seem to have been adequately handled and this fact tends to minimize the value of the findings.

Hartlage and Lucas (1973) used 1132 children as subjects in the validation of an approach to group screening for reading disabilities in the first grade. A correlation coefficient of .83 was achieved between teacher rankings of the children and the reading levels achieved by the students on the Wide Range Achievement Test. For comparison with the WRAT, 2 teachers' rankings were used. The result here, reflecting good levels of accuracy, was used to suggest that the trained observer, familiar with his subjects and the concepts under observation, can be considered likely to be quite accurate in his observations. The apparent thoroughness of this study gives its conclusions a good deal of merit.

Using a scale developed to measure 11 behavioral attributes, Lambert and Hartsough (1973) correlated multiple-teacher judgments of pupil characteristics. It was found that these multiple judgments correlated between .70 (often sick or upset under stress) to 1.0 (fighting and quarreling). On the basis of these high correlations between teacher

judgments, Lambert and Hartsough suggested that teachers are quite able to perceive and isolate the behavioral attributes of their students. The multiple rater technique, if and when practical, appears to be a useful method to determine a measure of reliability of an instrument, although the matter of establishing validity may still be elusive. This study appears to have picked useful attributes for study without being overly restrictive or unduly open-ended.

In an attempt to predict potential learning problems in low Socio-Economic Status rural children, Lessler and Bridges (1973) found a correlation of .75 ($p. < .001$) between results on the California Achievement Test and teacher ratings of pupil performance. This fairly high level of predictability was not found on other measures used. It appeared that the Metropolitan Readiness Test was the best predictor of potential learning disabilities. This research, though lending no great strength to the argument for teacher rating of pupils' performance, at least suggests that, in some areas, teachers can predict future performance based on their observations.

Maguire's (1973) work showed no significant differences between child care workers' ratings on an abridged Devereux Adolescent Behavior Scale and self ratings made by female adolescents with behavior disorders, on themselves, using the Teen-Agers Self Awareness Test. This work, though not conducted in an educational setting as such, suggested

that an observant individual, using a reasonably objective instrument could be expected to accurately detect and evaluate the behavior patterns of other individuals.

In a study of slum pre-schoolers, Richards (1973) found a moderate but significant product-moment correlation between teacher ratings and the Peabody Picture Vocabulary Test I.Q's. It is possible that a higher coefficient of correlation may have been achieved had the teachers had more experience with the children tested in the study. Nonetheless, it is of interest that some correlation exists, even at this early level, between a teacher's view of a child and an objective instrument's evaluation of his intelligence.

Richmond and Dalton (1973), in a study of 9-15 year old retarded students using the Coopersmith Self-Esteem Inventory, found that the child's self image, as reported on the Inventory was positively related to teacher evaluations of academic ability while teachers' ratings of social and emotional behavior could not be shown to correlate significantly. While the inability of the teachers' rating of social and emotional behavior to correlate with the students' rating of selves is discouraging, this study does have value in showing that teachers do appear sensitive to their students' behaviors in some ways. It would appear that a student's self image and the behavior affected by it, could relate to his academic performance to some extent.

This study also forces recognition of the fact that teachers, as observers, are far from infallible and are very likely most sensitive to achievement related behaviors.

Investigating the accuracy of teacher predictions on learning performance, Wang (1973) had 2 teachers estimate the Primary Education Project results of their classes which were made up of 12, 4 year olds and 13 kindergartners. The first teacher had a mean accuracy of 67.7% (50.0% to 88.9%), while the second had a mean accuracy of 76.2% (65.7% to 88.9%). These results, though perhaps not as accurate as anticipated by Wang, were significantly different from chance values. The implication here is that teachers' predictions, though not 100% accurate or totally consistent, do reflect a certain level of awareness of their students' characteristics and capabilities.

The available evidence points toward a structured rating of behavior as being a potentially reliable and valid technique of behavioral observation. Structured rating allows for consistency and frees the observer, or should free the observer, from making value judgments. Although free observation, by teachers and others, has moderate apparent success the more structured form of rating appears to be more effective.

Great claims have been made for some of the methods described above. However, most of the research above has not been replicated in any way, so the validity of the

results is in question. Also, many of the rating scales devised by authors have not been thoroughly validated, thus casting doubts upon exactly what they measure. Despite these limitations and those discussed above, there does seem to be a place for teacher observation of student characteristics and attributes. It is in the realm of revalidation and replication that further research is indicated in much of the work discussed above.

Ratings and the Criteria of Ratings

Swift and Spivak (1969) acquired 298 ratings of fifth grade achievers and underachievers. The achievement criteria used were subtest scores on a group test and teacher assigned report card marks. An analysis of the relationship between classroom behavior and the achievement criteria indicated that when a child was underachieving the fact was evident in both grade or test scores and general functioning in the classroom. Underachievers, it was shown, were clearly different from achievers in manifestations of overt maladaptive behaviors. The authors pointed out that the findings were particularly true when the achievement criteria used was the teacher's judgment of the quality of the child's efforts. This would suggest that there may be a relationship between the criteria of achievement and the rating of behavior. The objective criteria used showed similarity to the teacher's subjective grading, but it is left to

conjecture whether the rating based on subjectively graded achievement reflects behavior or whether the behavior resulted from expectation.

The key point to be made here is that results seem to vary based on criteria and that subjective criteria e.g. grades, may be biased or biasing. The more objective the criteria, the more useful the results.

Factors Relating to Raters

In addition to concerns about the efficacy of behavior rating, rating scales and criteria, questions arise about factors affecting the rater. Such questions as the rater's attitudes towards the subject being rated and sex differences among raters deserve some consideration and will be dealt with briefly.

Both Grgin (1969) and Walker (1970), in his initial work on the WPBIC, suggest that no significant sex differences appear, on the part of teachers, in the rating of pupil knowledge or behavior. Both authors indicate that in using rating categories and exercising rigor in their ratings, male and female teachers show no real differences. The possible bias, either for or against same sex subjects, does not prove to be a relevant factor in rating by teachers.

Also, representative of work examining the relationship between teachers' attitudes and their ratings of pupil behavior is Willborn's (1972) study comparing teachers'

scores on the Minnesota Teacher Attitude inventory to their rating of pupils on the Behavior Maturity Scale. From this work, it appears that a teacher's attitudes have little or no bearing on his ability to rate students objectively.

To summarize briefly, there is a pattern of evidence, though perhaps not conclusive, suggesting that teachers are capable of observing and recording the behavior of their students, regardless of their own sex or attitudes toward the students. These observations may be more useful if a structured, objective instrument is used.

III RATIONALE

Based on the discussion presented in the review of literature relating to ratings and rating scales, there appears to be considerable scope for follow up research on these types of instruments, since many scales used, including the Walker Problem Behavior Identification Checklist (WPBIC) have not been subjected to any form of subsequent study.

Given Walker's initial research, which appears thorough, it was decided to study the checklist in terms of one form of validity. The re-estimation of contrasted groups validity was thus chosen as the focus of study. The method used was based on Winer (1962, pp. 89-92). The reason for this form of study was the existence of two groups of potential subjects with relatively well known characteristics. A group of subjects identified as behaviorally disturbed was available and in the same schools other children were available who were not so identified. It was felt that if the instrument could distinguish between these two groups, the desired estimate of validity would have been achieved.

With the above in mind, the subjects were selected and rated. This rating yielded full scale scores and scores on 5 factors used on the scale. These scores were then compared and distributed to determine where differences occurred, the direction of differences and any potential

weaknesses in the discriminatory powers of the instrument.

To rule out sex differences in the scores, two subgroups were formed which were matched for age and sex and compared. These two groups were drawn from the original groups used in the study.

To summarize, additional research on the WPBIC appeared warranted and the estimation of contrasted groups validity appeared to be the most fruitful area of study in terms of the instrument's probable future use as an aid in identifying behaviorally disturbed children. Various approaches to this estimation were decided upon to obtain a relatively clear picture of the instrument's discriminatory properties.

Definition of Terms

1. Problem Behavior. Problem behavior will be operationally defined in terms of the specific behavior listed in the Walker Problem Behavior Identification Checklist (WPBIC) which comprise a set of 50 behaviors considered by raters and judges to be problematic, the list being drawn from observational reports by classroom teachers. Such behaviors would be broadly classified as acting out, withdrawal, distractability, disturbed peer relations, or immaturity. These broad categories are

purported to cover the 50 behaviors included in the Checklist.

2. Junior Adaptation Class. A Junior Adaptation class is one including children who are of normal intelligence, at least two years behind their peers in academic progress, and displaying behavior problems. These children are selected on the basis of extensive psychological evaluation, intelligence assessment and reports based on teacher observation over an extended period of time.

3. Acting-Out. Acting out will be considered to be any behavior which indicates defiance to, or outright refusal to, comply with teacher instructions. If a child overtly refuses, by statement or action, or both to carry out the teacher's instructions within a certain specified period of time, he would be considered to be defying the teacher. This type of behavior could also include argumentativeness, extreme affect in the face of frustration, overly aggressive acts, temper tantrums, distortion of the truth and undue approval seeking for tasks completed.

4. Withdrawal. Withdrawal will be defined as the absence of engaging in, initiating or responding to interactions with other children, whether of the same or opposite sex. The withdrawn student will also be considered the one who seeks not to draw any attention, either by the teacher or other students, to himself.

5. Distractability. Distractability will be defined as the inability to attend to task. The child who is considered distractable, will be the child who can be distracted from task by small movements and noises, who seems to be staring into space for long intervals, who underachieves and does not complete tasks consistently, or is overly meticulous, who tends to regularly disturb other children engaged in on task performance, or who seems unable to stay on task or within limits unless external control is applied.

6. Disturbed Peer Relations. The child whose peer relations are disturbed will be defined as a child whose relations are entirely with same sex children, who stammers or stutters and appears unable to communicate effectively with peers, who comments that no one likes him, yet will not allow well done work to be displayed, or who often mutters unintelligibly to himself, rather than communicate with others.

7. Immaturity. The immature child will display certain age inappropriate behaviors such as enuresis, nervous tics, excessive nail biting, psychosomatic reactions to stress, listlessness, or tiredness. The immature child may also be shunned or avoided by others because of his age inappropriate behavior and may chose younger children as his playmates since their interests and activities more closely approximate his own than do those of his peers.

8. "On-Task" and "Off-Task" Performance. These terms will be operationally defined as performance, either appropriate to the immediate learning situation as structured by the teacher (on-task) or inappropriate to that situation (off-task). In a broader sense, these terms can also be applied to situation appropriate or situation inappropriate behavior in a social context.

9. Misclassification. For the purpose of this study, misclassification will be taken to mean assignment of a score above the critical or cut off score on a factor or full scale to a regular class student, or below the critical or cut-off score for an Adaptation class student. This term does not imply that an error has necessarily been made in the identification of any particular subject.

Hypotheses

1. The mean overall checklist scores and variances will show significant differences between the regular class students and those in the Junior Adaptation classes. Differences will be sought beyond the .01 level of confidence. The two groups of subjects will be shown to be heterogeneous in terms of both mean scores and variances.

2. The mean checklist score on each factor will show significant difference between the regular class and those in the Junior Adaptation classes. Differences will be sought

beyond the .01 level of confidence.

3. The two groups will not overlap in scores achieved on the entire checklist beyond 15% of each group. Since Walker (1970) indicated that 10 to 20% of school children have behavior disorders, a percentage between these values was chosen as the acceptable level of overlap around the cut-off point between problem and non-problem behavior. If greater than 15% of the regular class subjects are over the cut-off point, and greater than 15% of the Adaptation class group are below the cut-off point, certain questions regarding the instrument's usefulness could be raised.

4. The two groups will not overlap in scores achieved on each factor behond 15% of each group.

5. Groups, matched for age and sex, will show significantly different overall checklist mean scores. Differences beyond the .01 level of confidence will be sought.

6. Groups matched for age and sex will show significantly different individual factor mean scores.

Design

The design of the study became fairly complex due to the nature of the selection of subjects for the study. The subjects were drawn from 5 schools in the local area. In each school two classes had members drawn from them for inclusion in Group I and two classes or major parts of classes were used to form Group II. An illustration of the

design of the groups for one school is presented in Figure 1.

Because of the design, a number of possible sources of variance emerged which required separate analysis. This analysis was performed "post hoc" and will be discussed later under Analysis Techniques.

It should be understood that, although this analysis was not performed as part of the original hypotheses, its use was necessary to determine the extent to which factors other than group differences afffected the checklist scores.

Illustration of Group Structure for 1 School

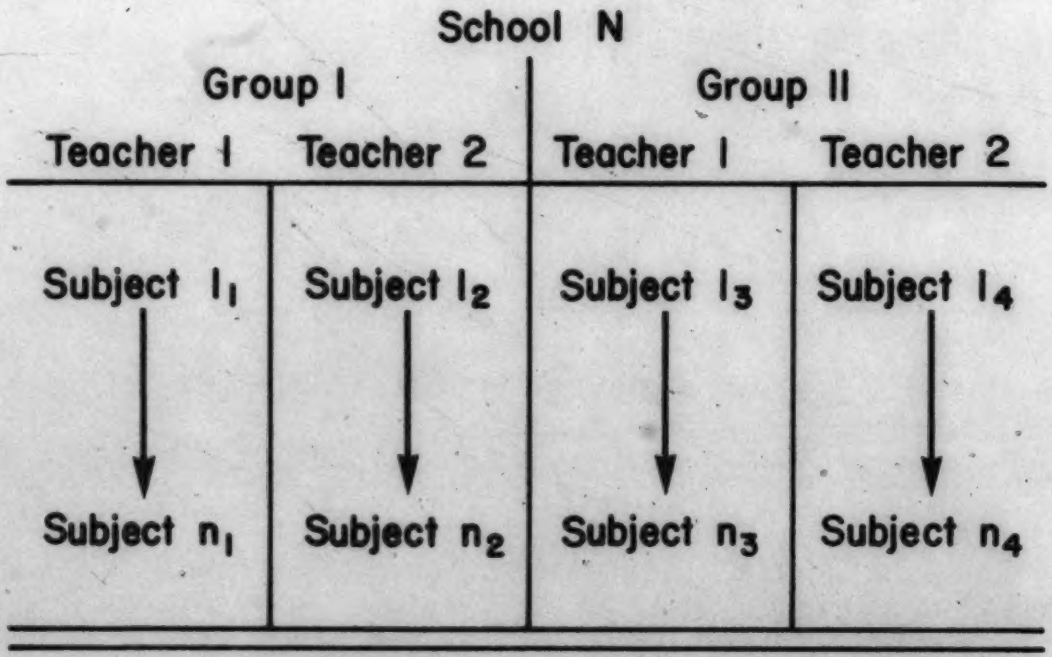


Figure I

IV METHODS

Subjects

The subjects, who ranged in age from 8 to 13 years, with a mean age of 10 years, 1 month, comprised a total of 188 students (126 male, 62 female) in two groups. Group I was made up of 94 students (47 male, 47 female) in 10 classes in 5 schools. Group II was made up of 94 students (79 male, 15 female) in 10 Junior Adaptation classes in the same 5 schools. The matching of the number of subjects from each school was done in an attempt to minimize the differences, between the groups, that might be attributable to differing school environments. It was hoped that this equality of numbers from each school, across the groups, would hold the school environment factor fairly constant.

Since the Adaptation subjects (Group II) consisted of whole classes, and due to the nature of obtaining permission to use classes for study purposes, true randomness of sampling was not possible for this group. This fact is noted as a limitation of the study. The regular class (Group I) subjects were drawn randomly from their classes.

Students from Junior Adaptation classes were selected to form one group of subjects because they satisfied at least one criterion for behaviorally disturbed children as outlined by Walker (1970). The class in which they have been placed

represents a special educational provision related to their problem behavior.

No socio-economic data, as such, were gathered on these subjects and only a cursory examination was made of their records. This examination suggested that the subjects, generally, were well distributed through the society in terms of socio-economic status. Further examination may, or may not, have shown a socio-economic status factor to be present as an extraneous variable, although this appeared unlikely. Some mention has been made of the remediation of disadvantaged, low S.E.S. children (e.g. Sibley, Abbott and Cooper (1969)) and that attitudes toward negative behaviors vary with S.E.S. (e.g. Piliavin and Briar (1964)), but other research has tended to support the concept that negative behaviors are not confined to status limits. A case in point was the aggression research in which aggressive behaviors were easily acquired by university undergraduates (see Berkowitz (1966)). Furthermore, Eron, Huesmann, Lefkowitz and Walder (1972) found that socio-economic status was less of a factor in one form of behavior problem (aggression) than were television viewing habits.

The Instrument

This study will examine the Walker Problem Behavior Identification Checklist (1970). This checklist was standardized using 534 Grades 4, 5 and 6 students in 21

classes in the Northwestern United States.

The checklist was made up of 50 operational statements, which were selected as the most frequently made of 300 statements submitted by teachers about problem behaviors. The statements were then weighted for severity by a group of judges with weightings ranging from 1 for the least severe to 4 for those items considered to reflect the most serious problem behaviors.

Reliability was estimated using the Kuder-Richardson split-half method which yielded a split-half correlation of .98. This correlation, according to Lindquist (1950), makes individual separations among subjects possible.

Four estimations of validity were obtained by Walker, including contrasted groups, criterion validity, factorial and item validity.

To establish contrasted groups validity, 38 children, meeting one of three criteria (a) psychological, psychiatric or clinical examinations; (b) specific educational provisions; (c) home instruction due to inability to benefit from school instruction—were matched with 38 students, not so identified, in terms of age, sex, and grade. Differences, significant beyond the .001 level of confidence were found between the means of the two groups. Walker thus claimed contrasted groups validity for the checklist.

Criterion validity was estimated using a biserial

correlation to measure the relationship between checklist scores and the construct of problem behavior as measured by the three criteria listed above. From this biserial correlation of .68 (standard error .039 and index of predictive efficiency of .33) Walker claimed that the instrument was useful in predicting behavioral disturbance at the elementary level.

A complicated procedure involving factor analysis was claimed to yield five, relatively independent, factors; namely Acting-out, Withdrawal, Distractability, Disturbed Peer Relations and Immaturity. Only Acting-out and Distractability overlapped significantly, thus intimating some common variance here.

Item validity indices were obtained for all checklist items which varied from .03 to .67. According to Walker, these indicated a high correlation with the total score and that the items discriminated between the upper and lower 27 percent of the sample in terms of scores. He also claimed that all but three items, numbers 33, 36 and 47, appeared to indicate a relatively homogeneous set of behavior.

Walker also found that boys scored significantly higher than girls on the checklist. This result may appear surprising, but Bandura (1973) points out, in specific reference to aggressive behavior that:

> . . . boys, who are generally encouraged to emulate feats of physical powers, spontaneously performed all they had learned when they saw aggression well received . . . . By contrast, girls, for whom physical aggression is traditionally regarded as sex inappropriate and, hence, negatively sanctioned, kept much of what they had learned to themselves . . . one should be more concerned with predisposing conditions than with predisposed individuals (p. 67).

The broad pattern of traditional socialization of children places physically aggressive and active roles within the expected roles of boys, not girls (see Lefkowitz, Walder, Eron and Huesmann (1973). It seems safe to assume, then, that boys will show more types of overt behaviors and problems, especially related to aggressiveness, than will girls. The differing social expectations could partly account for the greater number of general, overt, behavior problems among boys. It was also found, by Walker, that no significant differences appeared as the result of sex differences on the part of the rater.

Thus far, Walker's claims for the checklist have been the only views presented. In a critical review of behavior rating scales, Spivak and Swift (1973) present the following conclusions regarding the WPBIC:

> As an initial screening device, the WPBIC appears easy to use, probably taking no more than 5 minutes . . . Data suggest that the scale is reasonably homogeneous. Validity data, however, remain limited: the WPBIC total score correlated with the single clinical criterion, but this was probably due to the fact that the behaviors employed in defining the criterion measure were the same as those comprising the WPBIC.

> Unfortunately, data regarding validity and reliability of factor scores are not available, and, in at least two instances misleading labels are assigned to factors . . . it is impossible to determine how the weighting process built into the scoring system may affect the levels of validity when scores are tested against a variety of criteria (p. 64.

The criticisms expressed by Spivak and Swift (1973), above, are well founded. This study will not, necessarily, correct the problems, or answer the questions raised by these authors. It will attempt to determine if the instrument can discriminate effectively and meaningfully, between two supposedly different groups of children. Questions beyond that framework will be left for future research.

Procedures

According to the guidelines established by Walker (1970), a minimum two month observation period was set as a requirement for the teachers who rated the students on the checklist in order to ensure rater familiarity with the subjects. Raters were instructed to include only behaviors manifested during the two month observation period. All teachers who were involved in the study were given the same instructions for completing the checklists. These instructions included a review of the instructions displayed on the checklist (see Appendix II), and personal instructions on the degree of value judgment allowable. It was indicated, and stressed, that if a behavior described on the checklist

had been observed, during the observation period, it was to be recorded regardless of frequency.

(a) Raters. The subjects were rated by their regular teachers after the two month observation period. Instructions to each rater, as noted above, were the same in form and content. All raters were volunteers who agreed to assist in the study. Because of the voluntary nature of rater participation, these people may not have been totally representative of all teachers in the local school system. The assumption is made, however, that they are similar to local teachers generally.

(b) Data Collection. The data used for analysis were collected from teacher ratings of pupil behavior. These ratings were used due to the fairly comprehensive knowledge that these teachers possessed regarding their students' behavior and because teachers, generally, have been shown to be reasonably accurate, according to the literature, in their perceptions of student behavior. Outside observers were not employed because their observation period would necessarily have been less than the continuous two month period and this fact could have seriously confounded any results obtained. Furthermore, the instrument was designed as a device to be used by teachers, and outside observers would have rendered the results meaningless within the framework in which the instrument was designed.

Checklists were in the schools for an average period of just over one week and, when completed, were collected for analysis.

Data Analysis Techniques

1. Preliminary Analysis. In addition to analyses performed to test the hypotheses, a hierachal analysis was performed to determine the influence of nested variables. Each group from each school was examined with a total of 140 subjects from the original groups examined, (10 groups of 7 from each of the larger groups), in order to determine the extent of a nested school variable and the extent of a nested teacher variable. The procedure used was similar to that described by Winer (1962) involving analysis of variance of each individual class grouping, each school grouping (2 classes combined), each of the two main groups and the determination of error variance which was used as the basis fro comparison. In this way, the confounding effects of schools and teachers were determined as well as the treatment effect (extent of real group difference) and the variance attributable to error.

2. Hypothesis #1. The means and standard deviations (for calculating variance) were computed and compared for each group. The means were compared using a two-tailed t-test between means of independent samples while an F-test for differences between variances was used to determine variance differences. As noted in the hypothesis, differences

were sought beyond the .01 level of confidence.

3. Hypothesis #2. The mean on each factor was computed for each group and compared using a t-test between means of independent samples. Differences, again, were sought beyond the .01 level of confidence.

4. Hypothesis #3. A frequency polygon was plotted showing the number of subjects in each group at the various score values (0 recorded, 1 and following scores recorded in intervals of 4). The critical area of overlap was deemed to be at score value 21 (T score 60) which Walker (1970) indicated was the point allegedly separating normal from problem behavior.

5. Hypothesis #4. Frequency polygons were plotted on each of the sets of scores on each of the five factors. The critical score in the separation of problem from non-problem bahvior, on each factor, is as follows:

(1) Acting out	(Scale I)	Between 7 and 8
(2) Withdrawal	(Scale II)	5
(3) Distractability	(Scale III)	6
(4) Disturbed Pear Relations	(Scale IV)	3
(5) Immaturity	(Scale V)	Between 2 and 3

6. Hypothesis #5. The 47 male subjects in Group I were matched for age with 47 randomly selected boys in Group II (selection being random within age limits). These subjects

ranged in age from 8 to 12 years of age. The means for each group were computed and differences noted. As previously, differences significant beyond the .01 level of confidence were sought. The analysis was to have been performed using female subjects, but with only 15 females in Group II, it was felt that insufficient numbers, here, would not yield meaningful results.

7. Hypothesis #6. A procedure identical to that for hypothesis #2 and #5 was performed on the individual factors. Differences were sought beyond the .01 level of confidence.

8. Post Hoc Analyses

(a) Internal Consistency. The internal consistency of each of the two groups, and the two groups combined, was estimated using the Kuder Richardson 20 correlation method (see Ferguson (1971) p. 367). This method yielded an overall estimate of the degree of internal consistency within the groups and with groups considered as one.

(b) Matched Group Comparisons with Walker's (1970) Group. Group II subjects were compared with a group of subjects used by Walker (1970) to estimate contrasted groups validity The 94 group II subjects and 38 subjects identified by Walker as behaviorally disturbed were compared using a two-tailed t-test in the comparison of means and an F-test in

the comparison of variance. This was done to determine if any similarity in mean scores for these two groups existed and if they could be considered as homogeneous in terms of variance.

(c) Comparison of Matched Groups to Source Groups. Using a t-test, the two sub-groups, matched for age and sex, were compared to their source groups. Means were compared to determine if the sub-groups were similar to the groups from which they were drawn. Obviously these groups should have been representative to some extent and this analysis was performed to determine if that representativeness did, in fact, exist.

(d) Correlation of Group II Scores to Age and Length of Time in Adaptation Classes.

Age and length of time in Adaptation Classes were correlated for Group II. This correlation was computed to determine the relationship, if any, of these two factors to overall checklist scores. Additionally, this analysis could, it was felt, give some indication as to the effectiveness of the Adaptation program.

V RESULTS

Preliminary Analysis

Hierarchal Analyses. As a result of the rather complex structure and nature of the groups involved in this study, a number of potential sources of variance emerged. These sources of variance included variance due to real differences between the groups, variance due to school differences, variance due to differences in treatment or approach within the schools, variance due to teachers in schools being involved with different groups and variance within groups (the measure used as a basis for comparison).

A Hierarchal analysis was conducted on the full checklist and on each of the five sub-scales, following a method similar to Winer (1962).

The results of the Hierarchal analysis of the full scale are presented in Table 1.

In this analysis, 2 significant (beyond the .01 level of confidence) sources of variance emerged. A moderately significant teacher within treatment by schools effect was noted. This effect would seem to indicate that some of the variance in checklist scores was due to variations in the teachers dealing with the children in the two different settings, i.e., Adaptation class vs regular class. The reasons for this will be dealt with in the discussion. The teacher

Table 1

Hierarchal Analysis - Full Scale

Source of Variance	Sum of Squares	Degree of Freedom	Mean Squares	F	Critical Ratio (.01)
Treatment	5136.45	1	5136.45	61.94*	6.84
School	1101.58	4	275.39	3.32	3.47
Teacher by School	689.67	4	172.42	2.08	3.47
Teacher within Treatment by School	2400.69	I0	240.07	2.89*	2.47
Within	9951.75	120	82.93		

* Significant beyond .01 level of confidence

within treatment by school effect is too large to totally ignore and must be noted as a potentially confounding effect.

In the analysis of variance on the full scale, the main source of variance is that resulting from what has been labelled the treatment effect. This effect is a reflection of differences between the two groups, since the two group variances are compared. The ratio of variance attributable to real differences compared to within groups variance, at 61.94 (c.r. 6.84) is significant well beyond the .01 level of confidence. This effect would, as a result, appear to overwhelm the significant though moderate teacher within treatment by school effect noted above.

The results of the hierarchal analysis of Scale I (Acting-Out) are presented below in Table 2. The only significant (beyond the .01 level of confidence) source of variance evident on Scale I was that attributable to differences between the groups of subjects. It appeared from these results, that a real difference existed between the groups on this factor.

Table 3 includes the results of the hierarchal analysis on Scale II (Withdrawal). As in Table 1, two sources of variance were noted. Also like the results found in Table 1, the significant sources of variance were the treatment (differences in class type) effect and the teacher within treatment by school effect. In this case,

Table 2

Hierarchal Analysis - Scale I

Source of Variance	Sum of Squares	Degrees of Freedom	Mean Squares	F	Critical Ratio (.01)
Treatment	1131.46	1	1131.46	39.64*	6.84
School	99.76	4	24.94	.87	3.47
Teacher by School	70.19	4	17.54	.61	3.47
Teacher within Treatment by School	457.42	10	45.74	1.61	2.47
Within	3425.14	120	28.54		

* Significant beyond .01 level of confidence

Table 3

Hierarchal Analysis - Scale II

Source of Variance	Sum of Squares	Degrees of Freedom	Mean Squares	F	Critical Ratio (.01)
Treatment	31.11	1	31.11	6.99*	6.84
School	53.50	4	13.88	3.00	3.47
Teacher by School	5.67	4	1.41	.32	3.47
Teacher within Treatment by School	117.57	10	11.75	2.64*	2.47
Within	533.43	120	4.45		

* Significant beyond .01 level of confidence

however, the teacher within treatment effect must be given considerable attention as the treatment effect was proportionately much less extensive here than in the full scale. It appears, that although the treatment effect appeared greater, it was seriously confounded by the teacher differences. The effect of the differences between classes (treatment) could best be described, here, as tentative because of the confounding teacher variance.

Table 4 shows the results of the hierarchal analysis of Scale III (Distractability). As in Table 1, a significant treatment effect occurred in regard to Scale III. Other potential sources of variance were not significant beyond .01 level of confidence. It appears that real differences, between the two groups, exist in terms of this scale. The rather pronounced F ratio between treatment effect and within groups effect suggests that in terms of this scale the groups are considerably different.

On Table 5 are presented the results of the hierarchal analysis performed on Scale IV (Disturbed Peer Relations). A significant treatment effect (beyond .01 level of confidences) was obtained on this scale and although not significant beyond the .01 level of confidence a moderate teacher effect was noted. The treatment effect appeared to be sufficiently powerful to account for the greatest amount of the variance on this scale.

Table 4

Hierarchal Analysis - Scale III

Source of Variance	Sum of Squares	Degrees of Freedom	Mean Squares	F	Critical Ratio
Treatment	244.47	1	244.47	34.82*	6.84
School	78.50	4	19.62	2.78	3.47
Teacher by School	80.92	4	20.23	2.88	3.47
Teacher within Treatment by School	145.78	10	14.58	2.07	2.47
Within	842.85	120	7.02		

* Significant beyond .01 level of confidence

Table 5

Hierarchal Analysis - Scale IV

Source of Variance	Sum of Squares	Degrees of Freedom	Mean Squares	F	Critical Ratio (.01)
Treatment	92.83	1	92.83	15.52*	6.84
School	22.52	4	5.62	.94	3.47
Teacher by School	27.10	4	6.77	1.13	3.47
Teacher within Treatment by School	142.99	10	14.30	2.39	2.47
Within	717.43	120	5.98		

* Significant beyond .01 level of confidence

Table 6 presents the final hierarchal analysis performed on the checklist and was performed on the fifth factor (Immaturity). As in the previous tables, the treatment effect appeared as the main source of variance, being significant beyond the .01 level of confidence. None of the other potential sources of variance differed significantly from the within groups variance measure. The difference between Groups I and II on the fifth factor (Immaturity) appeared to be considerable and real.

A treatment effect was noted on all five factors of the checklist as well as on the Full Scale. In none of the scales was there a significant (beyond .01 level of confidence) school or treatment within school effect. That is, when variances by schools were compared, and when variances by group within the schools were compared, there was not a significant effect on the overall variance of the groups. Moderate but significant teacher effects were noted on the full scale of the checklist and on Scale II(Withdrawal). Although this teacher effect did not seriously compromise the treatment effect on the full scale, because of the magnitude of the treatment effect, the same cannot be said for Scale II. The relative value of statements regarding this scale must be weighed against the effect of variations among teachers in rating subjects on this scale. It appears that very few definitive statements can be made regarding the discriminate powers of this scale vecause of the teacher effect.

Table 6

Hierarchal Analysis - Scale V

Source of Variance	Sum of Squares	Degrees of Freedom	Mean Squares	F	Critical Ratio (.01)
Treatment	51.60	1	51.60	13.54*	6.84
School	30.33	4	7.58	1.99	3.47
Teacher by School	36.21	4	9.05	2.37	3.47
Teacher within Treatment by School	143.36	10	4.34	1.14	2.47
Within	428.29	120	3.81		

* Significant beyond .01 level of confidence

1. Hypothesis #1. It was hypothesized that the means of the two groups currently under study would show differences beyond the .01 level of confidence. According to Ferguson (1971, p. 218), the performance of this analysis and the hierarchal analysis may have appeared redundant. This analysis encompassed all subjects under study, not merely a selected group, in order to yield an accurate picture of the extent of differences between the groups.

The mean for each group on the full scale was computed and compared using a t-test for means from independent samples. Table 7 (p. 56) presents the results of the

The full scale mean of Group II (Adaptation) was significantly different from that of Group I (regular), with differences significant beyond the .001 level of confidence. The mean of the 94 subjects in Group II appeared very similar to that of Walker's (1970) group of 38 experimental subjects used to determine contrasted groups validity. This group, like the Adaptation group in the current study, met certain criteria (see Walker (1970), p. 3) to differentiate them from the control group used for a comparison. It was this apparent similarity which prompted certain further analysis which will be discussed later.

Table 7

Comparison of Means

Scale	Group #	Mean	T-value (df = 186)	Critical Value
Full	I	4.95	-8.261***	+ 3.291
	II	16.71		-
I	I	1.28	-7.097***	+ 3.291
	II	6.73		-
II	I	.64	-2.583**	+ 2.576
	II	1.49		-
III	I	1.72	-6.970***	+ 3.291
	II	4.59		-
IV	I	.63	-4.346***	+ 3.291
	II	2.20		-
V	I	.67	-3.737***	+ 3.291
	II	1.70		-

*** Significant beyond .001 level of confidence
** Significant beyond .01 level of confidence

Table 8

Comparison of Variance Full Scale Only

Group	Standard Deviation (S)	Variance (S^2)	F	Critical Ratio $F_{(.99)}$ (93.93)
II	11.93	142.40	2.95**	1.59
I	6.95	48.27		

** Significant beyond .01 level of confidence

The variances were also compared, on the full scale, in order to determine the homogeneity or heterogeneity of the two groups. The data used and results obtained are presented on Table 8 (p. 57). This comparison was made using all the subjects, rather than the 140 subjects used in the hierarchal analysis to determine if differences in numbers substantially altered the variance.

The variances of the two groups showed a difference significant beyond the .01 level of confidence. These results. showing that the variance of Groups I and II were heterogeneous, and the apparent similarity of Walker's experimental group to Group II, prompted further investigation. The results of that investigation will be presented later.

2. Hypothesis #2. Table 7 (p. 56) presents the comparison of group means on the sub-scales of the WPBIC. All subjects were used in these comparisons.

On Scale I (Acting Out) differences between Groups I and II appeared which were significant beyond the .001 level of confidence. It was on this scale, measuring the extent of acting out behavior, that the greatest difference between the groups was noted.

The analysis of Scale II (Withdrawal) means showed that differences beyond the .01 level of confidence existed between the two groups. Caution in attempting to interpret

these results appears necessary because of the findings of the hierarchal analysis which will be presented later. Teacher effects confounded the results on this scale, so that, although differences beyond the stipulated level of confidence were found, these differences may have been caused by other than real differences between the groups.

Mean differences between groups on Scale III (Distractability) were found to be beyond the .001 level of confidence. Based on this finding it would appear that Group I was significantly less distractable than Group II.

Means on Scale IV (Disturbed Peer Relations), when compared, showed that Group II's mean score on this factor was significantly higher (beyond the .001 level of confidence) than that of Group I.

Differences beyond the .001 level of confidence were also found between the group means in Scale V (Immaturity).

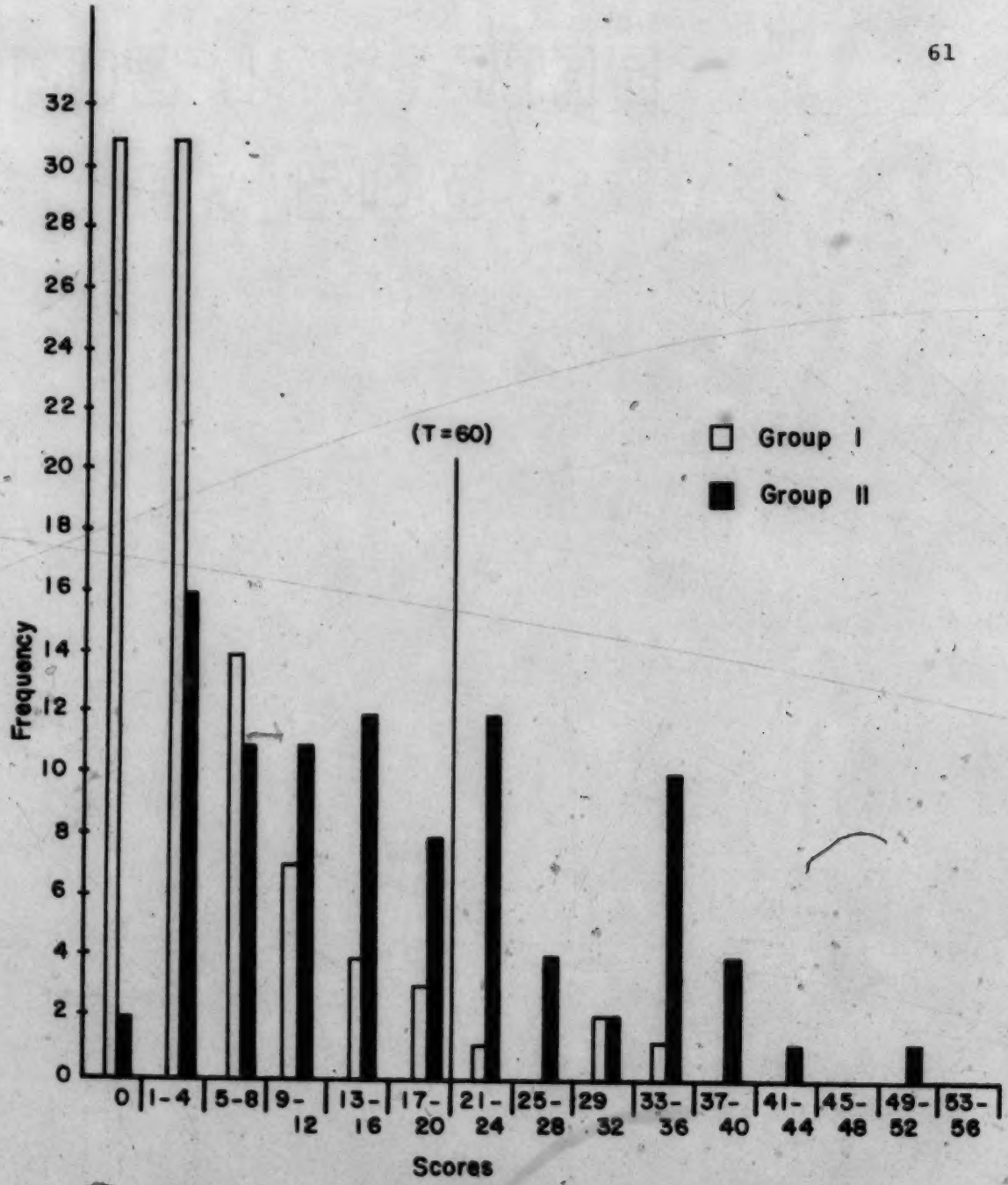
The comparison of means showed that on all five sub-scales, as well as on the full scale, Group II had higher mean scores than did Group I, and that the difference was, in all cases, significant beyond the .01 level of confidence. All but the Scale II differences were significant beyond the .001 level of confidence.

Scale II (Withdrawal), as the poorest indication of the scales presented here, still appeared to be a good indicator

of problem behavior along the dimension of withdrawal. However, as will be noted later, its ability to discriminate may have been weakened by outside factors and may not have been as adequate a scale as these initially presented data would suggest.

3. Hypothesis #3. The distribution of the 188 subjects on the full scale and sub-scales of the WPBIC were plotted and the amount of overlap of each group was shown. The percentage of misclassification was also calculated for each group. "Misclassification", in this case, was taken to mean the rating of Group I subjects above the respective critical scores as defined by Walker (1970) on the full scale and sub-scales, and the rating of Group II subjects below these critical points. This does not, necessarily imply that these subjects were erroneously rated.

The distribution of subjects on the Full Scale is shown on Figure 2. In order to keep this graph from becoming unduly awkward or crowded, all scores above 0 were grouped into intervals of 4. Therefore all subjects rated between 1 and 4, 5, and 8, etc. were grouped together. In Group I, 4 subjects were scored above a score of 21, the point established by Walker (1970) as the dividing point between problem and non-problem behavior. This represented 4.26% of Group I. 34 subjects, representing 36.17% of Group II were rated at or above 21. This left 63.83% of the group



Scores: Intervals above 0 represent intervals of 4 (1-4, 5-8, etc.)

Figure 2 - Distribution of Full Scale Scores

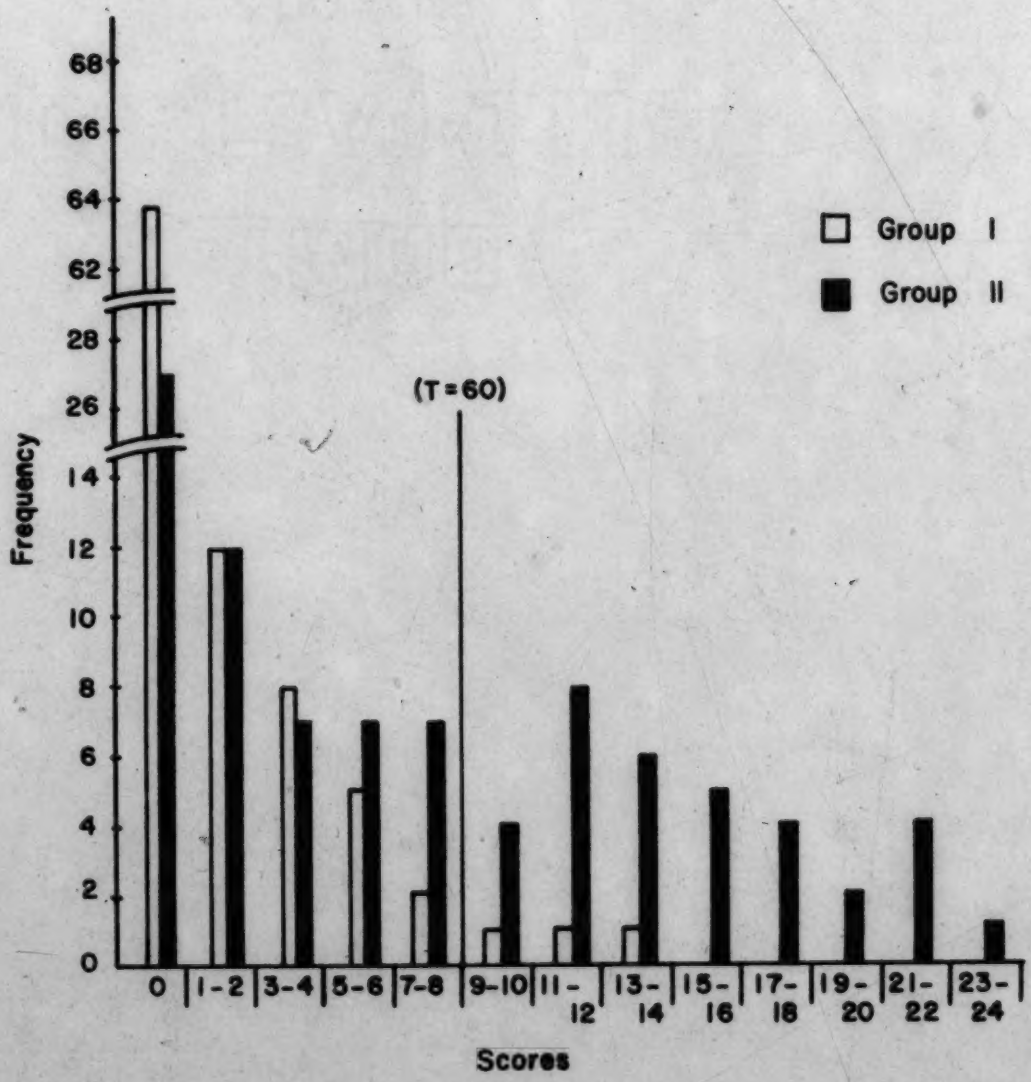
below this point.

4. Hypothesis #4. Figure 3 presents the distribution of subjects on Scale I. On this scale 4 subjects or 4.26% of Group I were rated above a score of 7, which is the critical score, according to Walker (1970). Of Group II, 39.36% (37 subjects) were above this point leaving 60.64% of the group below this score. On the graph used for Figure 3, each score is represented and the number of subjects at each recorded. This procedure was also followed on the remaining graphs.

Figure 4, presents the distribution of the subjects on Scale II (Withdrawal). On this scale 5.32% (5 subjects) in Group I were rated at or above the critical score of 5. Of the subjects in Group II, 14.89% (14 subjects) were scored at or above the critical score with 85.11% below this score. No clear discriminatory ability can be claimed here because of the confounding teacher effect which has affected the confidence placed in this scale on all of the analyses performed.

On Figure 5, 8.51% (8 subjects) of Group I were rated at or above the critical score of 6 while 37.23% (35 subjects) of Group II subjects were scored above the critical score on Scale III (Distractibility). This left 62.77% of Group II below this critical score.

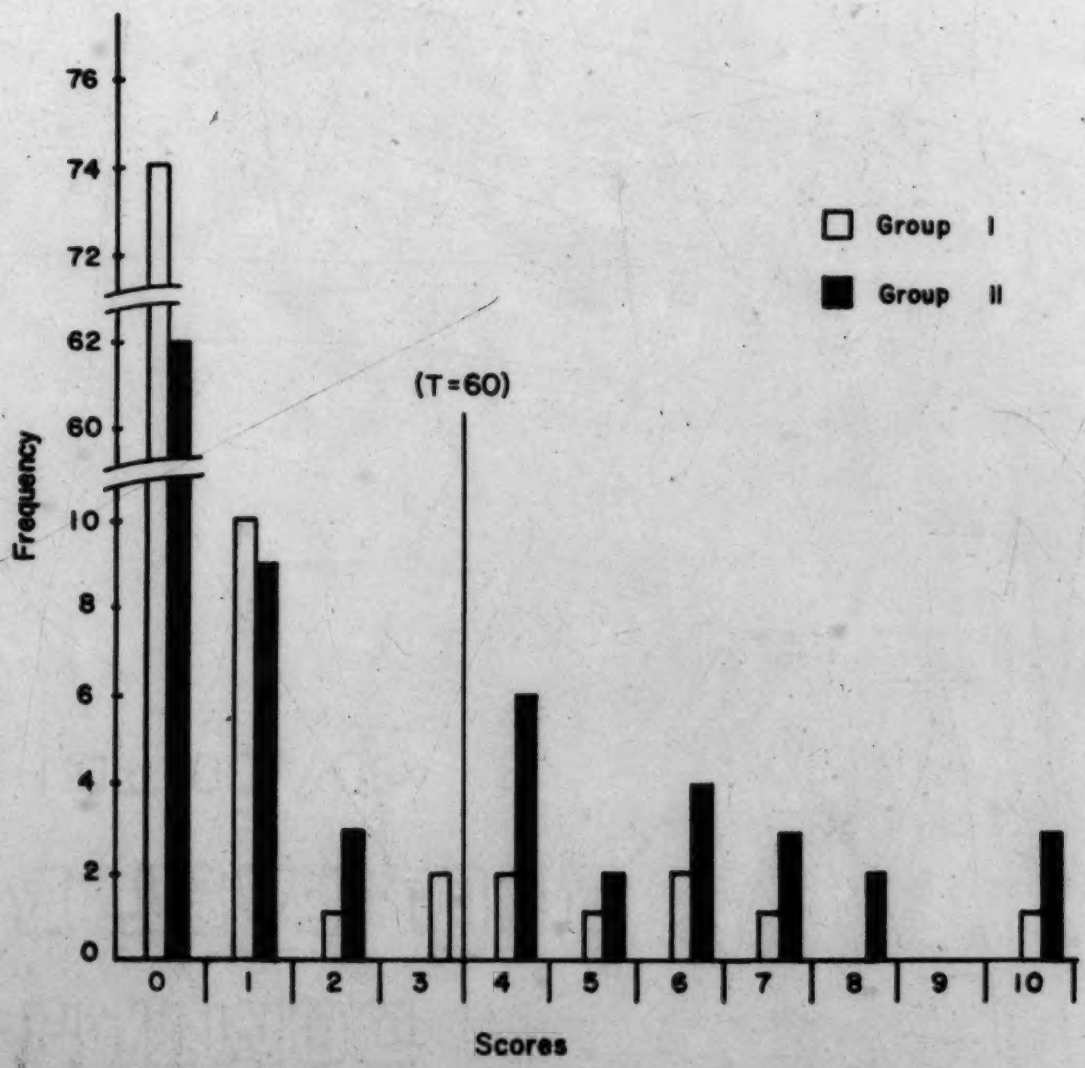
Figure 6, presenting the distributions on Scale IV,



Scores: Intervals above 0 represent intervals of 2

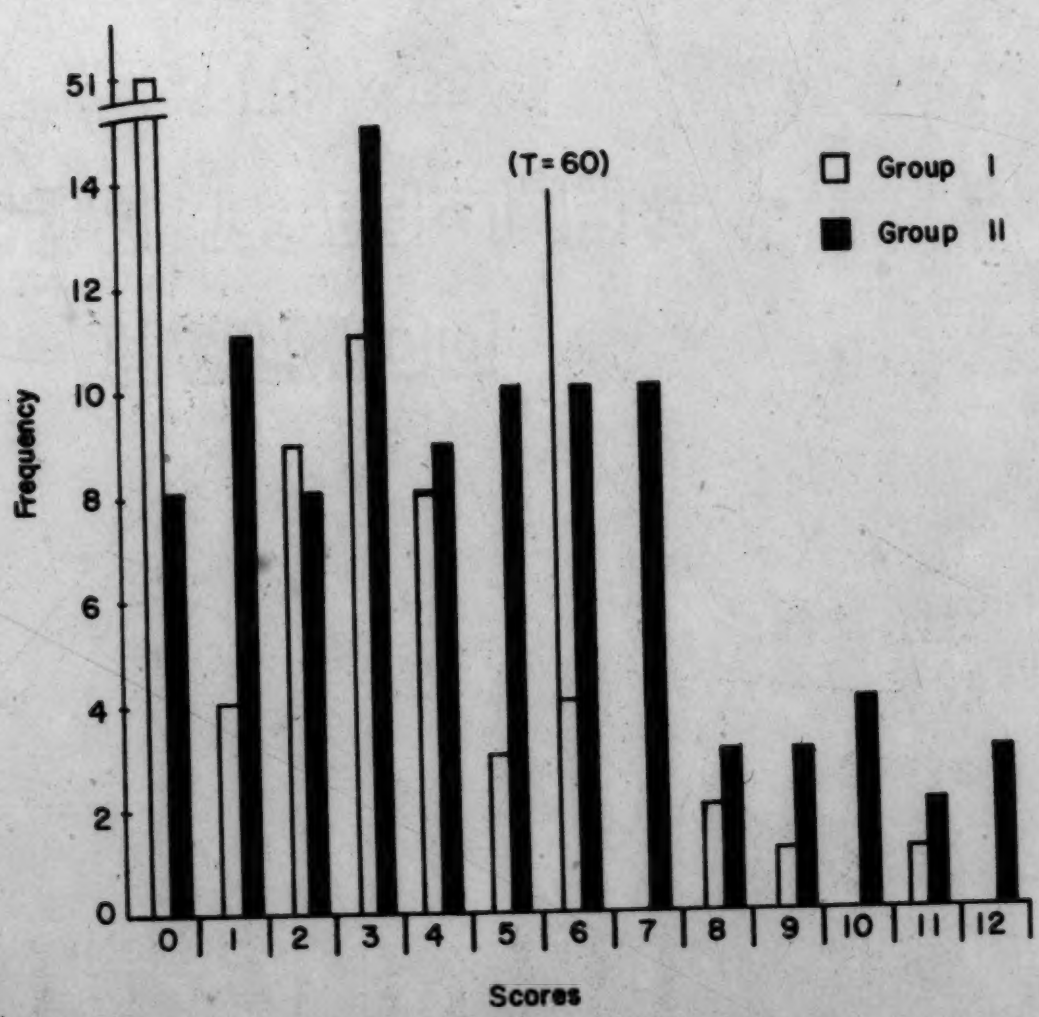
Frequency: Intervals of 2 with interval of 12 from 14 to 26 and interval of 34 from 28 to 62

Figure 3 - Distribution of Scale I Scores



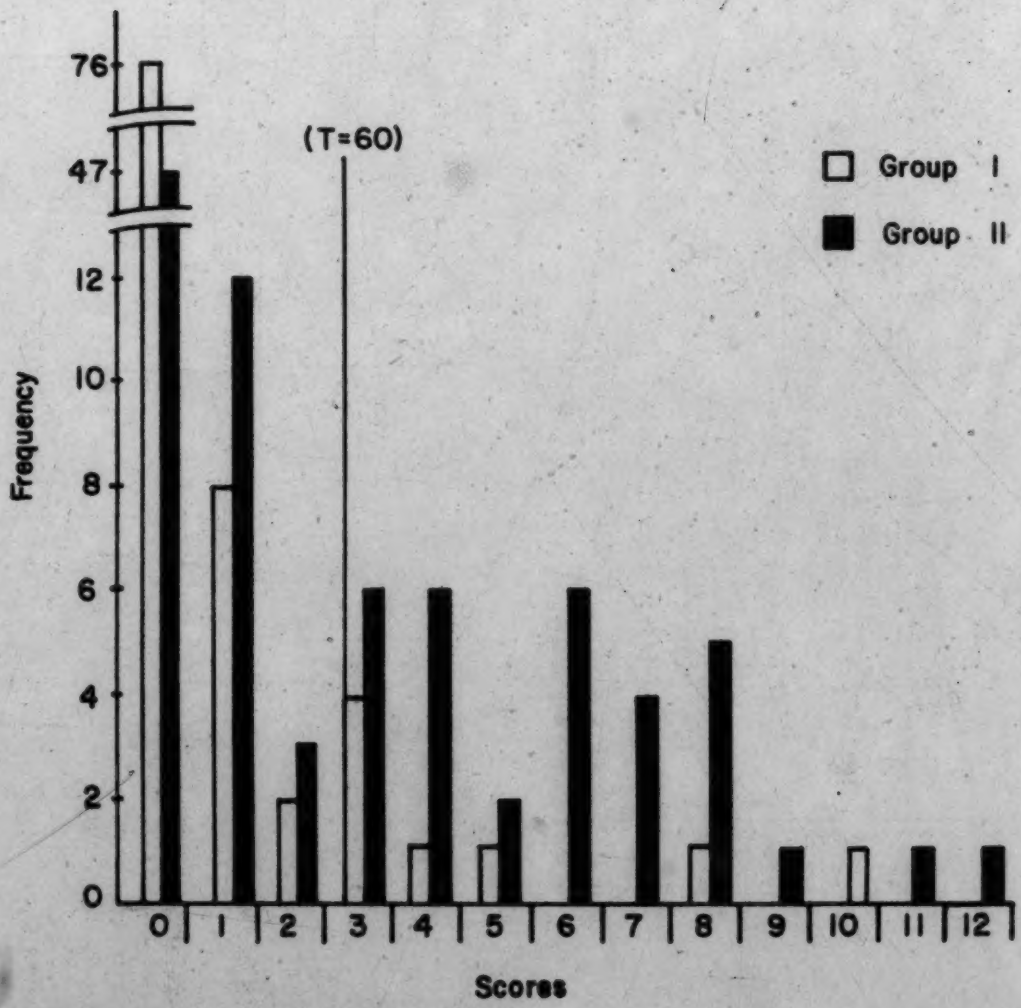
Frequency: Intervals of 2 with interval of 50 from 10 to 60 and interval of 10 from 62 to 72

Figure 4 - Distribution of Scale II Scores



Frequency: Intervals of 2 with interval of 37 from 14 to 51

Figure 5 - Distribution of Scale III Scores



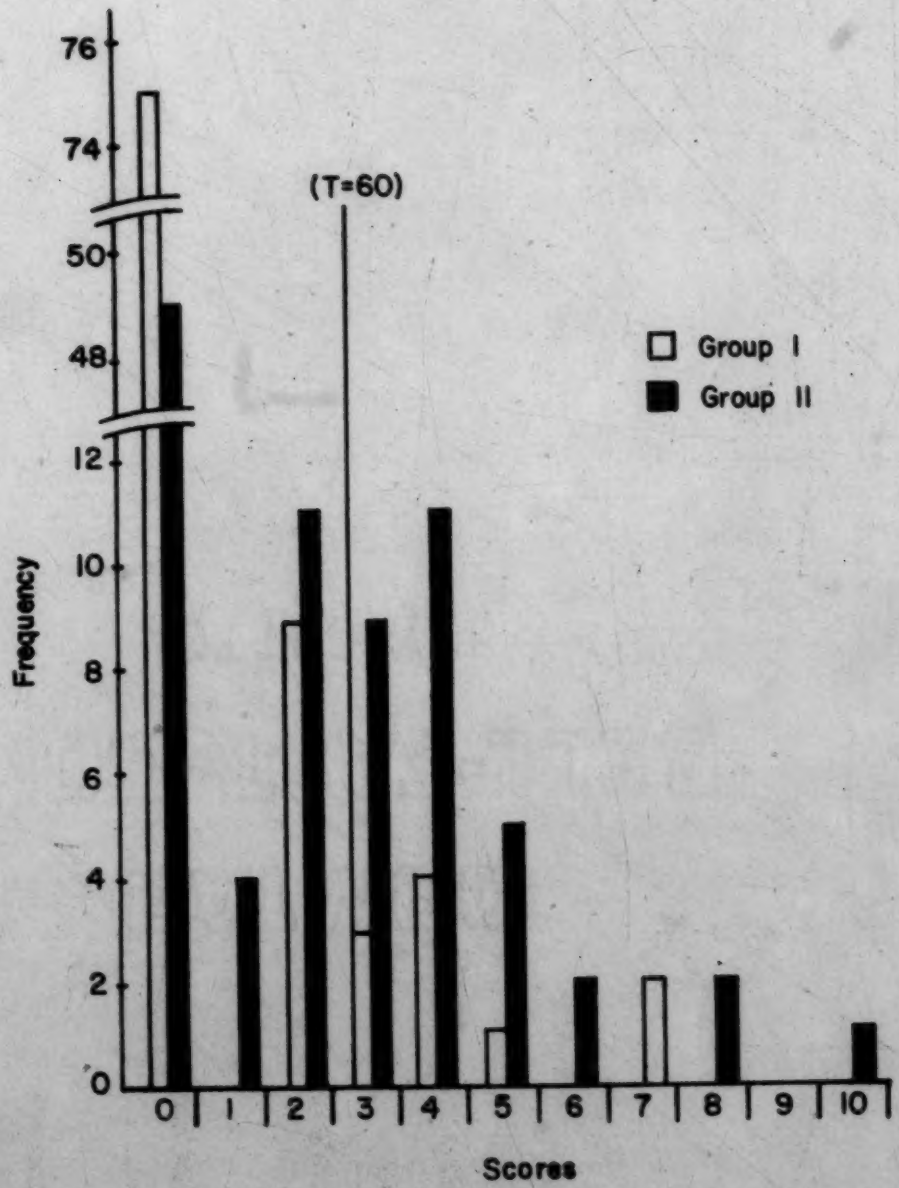
Frequency: Intervals of 2 with interval of 35 from 12 to 47 and interval of 29 from 47 to 76

Figure 6 - Distribution of Scale IV Scores

(Disturbed Peer Relations) shows 8.51% (8 subjects) of Group I at or above the critical score of 3. On this scale 34.04% (32 subjects) in Group II were at or above the critical score of 3 while 65.96% of the subjects were below this score.

Finally, as shown on Figure 7, 10.64% (10 subjects) of Group I scored above the critical score of 2 on Scale V (Immaturity). Of Group II, 31.91% (30 subjects) of Group II were rated above the critical score, while 68.09% of this group were rated below the critical score.

In looking at these graphical results, it can be noted that on each scale, the number of Group II subjects above the critical score substantially exceeded the number of Group I subjects. On Scale I (Acting Out) 9.25 times as many subjects in Group II were rated above the critical score as were Group I subjects. The ratios on the other score ranged between this ratio and a 2.79 Group II to Group I ratio on Scale II (Withdrawal), the weakest scale in terms of ability to discriminate between groups. The remaining Group II to Group I ratios, including that for the full scale were 8.5: 1 for the full scale 4.375: 1 for Scale III (Distractability), 3.99 for Scale IV (Disturbed Peer Relations), and 2.99: 1 for Scale V (Immaturity). The ratios presented above were, to reiterate, the ratios of Group II subjects rated at or above the critical score for a scale, to Group I subjects above that score.



Frequency: Intervals of 2 with interval of 36 from 12 to 48 and interval of 24 from 50 to 74

Figure 7 - Distribution of Scale V Scores

5. Hypothesis #5. From each group, 47 subjects were matched for age and sex. Male subjects ranging in age from 8 to 12 years were compared, with equal numbers from each group at each age level. The mean scores on the full scale were compared for statistical significance using a two-tailed t-test with differences sought beyond the .01 level of confidence. The results of this comparison are presented in Table 9.

Difference , significant beyond the .001 level of confidence, was found between the full scale means of the groups used in this analysis. These results showed a pattern similar to those achieved using the full groups although the matched group means were somewhat higher than those of the full groups.

6. Hypothesis #6. Using the scores of the subjects mentioned, above, the mean scores on each scale were compared for statistical significance beyond the .01 level of confidence. A two-tailed t-test was used as above. The results are presented in Table 9.

Differences, significant beyond the .001 level of confidence, were noted on the means of Scales I (Acting Out), III (Distractability) and IV (Disturbed Peer Relations). On Scale V (Immaturity), a difference between the matched group means was found which was significant beyond the .05 level of significance. This difference was below the confidence level

Table 9

Comparison of Means. 94 Subjects Matched for Age and Sex

Scale	Group #	Mean	T-value (df-92)	Critical Value
Full	I	6.26	-5.883***	± 3.417
	II	18.94		
I	I	1.81	-4.935***	± 3.417
	II	7.60		
II	I	.87	-.1418	± 1.99
	II	1.60		
III	I	2.21	-4.789***	± 3.417
	II	5.26		
IV	I	.62	-3.753***	± 3.417
	II	2.40		
V	I	.72	-2.153*	+ 1.99(.05)
	II	1.64		- 2.639(.01)

*** Significant beyond .001 level of confidence
* Significant beyond .05 level of confidence

sought. Scale II failed to discriminate between the groups even at the .05 level of confidence.

Post Hoc Analyses

(a) Internal Consistency. Separate estimates of reliability were obtained for each group. The results of a Kuder Richardson Split-Half reliability measure, similar to that performed by Walker, but where the relation between all possible pairs was obtained, yielded a reliability coefficient of .8117 for Group I (regular class) and .8143 for Group II. Also, a similar measure was obtained for the pooled groups to determine the level of consistency of the raters as a group. This calculation yielded an overall reliability of .8598. This was not an entirely satisfactory estimate of reliability as the 20 raters were considered as 2 raters when the estimates were made for the groups separately and as one rater when the general measure was taken across the groups.

The overall estimate of internal consistency across the groups suggested that approximately 86% of the variance obtained in the checklist scores was true variance while approximately 14% of the variance was due to error (see Ferguson (1971, p. 365). Walker (1970) cited Lindquist

(1950) in stating that:

> With a reliability coefficient of .98, the checklist is capable of making individual separations among subjects with a considerable degree of reliability as an r of .90 is the minimum coefficient acceptable for the purpose (p. 3).

However, Lindquist (1950) in discussing the levels of reliability based on the guidelines set by Kelley (1927) said:

> . . . making the assumption that for a test to be useful, it must permit discriminations of a difference as small as 0.26 times the standard deviation of a grade group with chances 5 to 1 of being correct. Kelley arrives at the following as the minimum correlation for several purposes.
> . . . (b) To evaluate differences in level of group accomplishment in two or more performances . . . . . .90.
> It must be recognized, however that these values are arbitrary, being derived from the above assumptions as to what would be reasonable to expect a test to do in the way of discriminations between individuals and groups (p. 609).

Walker (1970) achieved a reliability coefficient of .98 using 534 sets of observations. In examining each of the groups separately, a mean reliability of about .813 was estimated, while the combined group reliability coefficient was .8598, using all 188 observations. As the sample size increased, so did the estimate of reliability. As Ferguson (1971) points out:

> Low reliability does not necessarily invalidate a technique as a device for drawing valid inferences. Low reliability may be compensated for by increasing sample size . . . . When significant results are reported with an unreliable technique on a small sample, the treatment applied is usually exerting a gross effect (p. 373).

An overall reliability of .8598 cannot be considered especially low. Neither did it approach Walker's coefficient of .98. Observation of the trend of increase of the reliability coefficients from that of the two groups separately to that of the groups pooled, suggested that with an increase sample size, the coefficient of reliability would increase to a point approximating Walker's figure. The value of increasing the sample sizes for that purpose would have provided little real benefit in terms of increased reliability since the coefficient obtained was already of a fairly high order.

(b) Matched Group Comparisons with Walker's (1970) Experimental Group

The apparent similarity between Group II mean and that of Walker's (1970) experimental group of 38 behaviorally disturbed subjects prompted a statistical comparison of those two groups. The results of that comparison are presented below. A two-tailed t-test for differences of independent means was used following Ferguson (1971, p. 152).

As can be observed from Table 10, no significant difference appeared on this comparison. It appeared that Walker's (1970) group of experimental subjects was similar to the group of 94 subjects in the Adaptation class group used in the current study. No comparison of this sort was possible on the five sub-scales as those data were not presented in Walker (1970).

Table 10

Comparison of Means of 38 Subjects from Walker (1970) and 94 Junior Adaptation Students

Scale	Group	Mean	T.value (df = 130)	Critical Value
Full	Walker (1970) Exp.	16.63	.03	+ 1.960
	Adaptation	16.71		-

For comparative purposes, the variance of the Adaptation group in this study and that of Walker's (1970) experimental group were also examined for statistically significant differences. The results of that comparison are presented in Table 11. The difference in variance between the two groups were not statistically significant. As in the case of the means, Walker's (1970, p. 3) experimental group appeared quite similar to the Adaptation group used in the present study.

(c) Comparison of Matched Groups to Source Groups I and II

In order to determine if the two groups matched for age and sex were similar to or dissimilar to the groups from which they were drawn, the means on the full scale of the matched groups were compared to the various means of the respective groups from which they were drawn. As previously, a two-tailed t-test was used for comparison and the results are presented in Table 12.

The results presented on Table 12, indicated that the two groups which were matched showed no significant differences from the respective source groups. As noted previously in regard to the matched groups, the WPBIC could not distinguish between the groups in Scale II and could do so only at a low level of confidence (.05) on Scale V.

(d) Correlation of Group II Scores with Age and Length of Time in Adaptation Classes

A simple correlation was performed between checklist

Table 11

Comparison of Variances of 38 Subjects from Walker (1970) and 94 Junior Adaptation Students

Group	Standard Deviation (S)	Variance (S^2)	F Ratio	Critical Ratio (.05) F (.95) (37, 93)
Walker's Experimental	12.68	160.88	1.13	1.51
Adaptation	11.93	142.40		

Table 12

Comparison of Means - Matched Group Means to Source Group Means

Scale	Group #	Mean	T-value (df = 139)	Critical Value
Full	I - Full (N=94)	4.95	.956	2.576 (.01)
	I - Matched (N=47)	6.26		
Full	II - Full (N=94)	16.71	.842	2.576 (.01)
	II - Matched (N=47)	18.49		

** Significant beyond .01 level of confidence

total scores, age and the length of time each student in Group II had been in a Junior Adaptation class. The mean total checklist score was 16.702, the mean age was 128.574 months (standard deviation 13.356) and the mean length of time in Adaptation classes was 15.308 months (standard deviation 8.453).

The coefficients of correlation obtained showed tendencies toward negative correlations although they were not of sufficient magnitude to warrant a great deal of emphasis. A correlation coefficient of -.143 was obtained between the checklist total score and age, while a coefficient of -.054 was obtained between the checklist total score and length of time in the Adaptation program. These results would indicate that in the case of the present group under study, there is a minimal relationship between age and behavior as rated on the WPBIC and length of time in Adaptation classes and behavior as measured on the WPBIC.

To summarize briefly, it was found that Group I and II differed significantly on the full scale of the WPBIC and on all of its sub-scales. A variance estimate using all of both Groups revealed that their variances were not homogeneous.

The graphed distributions of the subjects showed substantial differences in numbers of subjects above the critical scores on all scales except, perhaps, on Scale II where the differences in numbers though present were not

clearly dissimilar. Only 14.89% of Group II scored above the critical score of 5, on Scale II (Withdrawal) while 5.32% of Group I were rated above this point. According to Walker (1970), between 10% and 20% of a regular group could be expected to have behavior disorders, and, if true, even Group II would have been similar to a regular group on this scale.

When 47 boys from each group were matched for age and compared on the full scale and all sub-scales, it was found that differences beyond the .001 level of confidence existed on all scales but Scale II and Scale V revealed differences significant beyond the .05 level of confidence; Scale II (Withdrawal) showed no significant differences between the matched groups. The groups, later determined to be generally similar to their source groups, generally followed the pattern of other analyses in that Scale II appeared to be the weakest scale in terms of ability to discriminate between groups. Scale V, with differences beyond the .05 level of confidence was also less able to discriminate between groups in this analysis than in other analyses.

It was determined that differences between Groups I and II account for the greatest amount of the variance on the full scale and Scales I, III, IV, and V. On the full scale and Scale II, a teacher rating effect beyond the .01

level of confidence was also observed. Because of the overwhelming effect of Group differences on the full scale, the teacher effect, which was significant but low by comparison, was de-emphasized. However, because of the only moderate significance of the group differences on Scale II, the teacher effect was considered to be of sufficient magnitude to compromise the value of that scale in determining the presence or absence of Withdrawal type behavior.

A measure of internal consistency was also estimated indicating that approximately 86% of the variance obtained in checklist scores was due to real variance while the remaining 14% could be attributed to error.

When the members of Group II were compared to Walker's 38 experimental subjects it was found that Group II and Walker's group had similar means and variances. Also, when the matched groups of the present study were compared to their source groups, they were found to be similar to the groups from which they were drawn.

A correlation between age, length of time in Adaptation classes and overall checklist scores in the case of Group II showed minor negative coefficients between scores and age and scores and length of time but these did not prove to be significant.

VI DISCUSSION AND CONCLUSIONS

Discussion

This examination of the Walker Problem Behavior Identification Checklist has revealed certain strengths and weaknesses in the checklist which will be discussed in some detail.

The comparisons of mean scores of the two groups suggest that real differences existed between the two groups of subjects in the present study. With mean score differences significant beyond the .001 level of confidence in the full scale as well as on Scales I (Acting-Out), III (Distractability), IV (Disturbed Peer Relations), and V (Immaturity) with the higher mean score always associated with Group II, it appears that the instrument has detected differences, between the Groups, in the nature of manifest behaviors, and was useful in detecting instances of Acting-Out, Distractability, Disturbed Peer Relations and Immaturity, as well as differences between the Groups in terms of overall behavior. A similar claim cannot be made for Scale II (Withdrawal) despite the fact that mean scores differed beyond the .01 level of confidence. Scale II, as noted earlier, was subject to a rather serious teacher effect which compromised much of the interpretative value of this scale.

With the above noted exception the instrument appeared quite strong in its ability to distinguish between

groups.

The analysis of variance using both full groups indicated that the two groups were not homogeneous and suggested the possibility that the groups could represent different populations.

From the comparisons of means and variances, the evidence tends to support hypotheses #1 and #2.

When the hypotheses for the results of graphing the scores were developed, it was stated that a "misclassification" of greater than 15% for either group would raise questions about the value of the instrument. Within that frame of reference the value of the instrument is in serious doubt because the percentage of misclassified subjects in Group II ranged from 60.64 per cent for Scale I and 85.11% for Scale II with an average of 67.73% of subjects being "misclassified." This problem did not arise with Group I where percentages of "misclassification" ranged from 4.26% for the Full Scale and Scale I to 10.64% for Scale V with an average of 6.92% of the subjects being "misclassified."

According to Walker (1970) up to 10 to 20% of students can be expected, in the regular class, to display serious behavior disorders, while Freehill (1973) states:

> From reviewing the data, White and Harris (1961) concluded that a figure for a mild disturbance was impractical but a working estimate for serious maladjustment was between 2 and 12 percent. A practical and widely used rate comes from Bowen (1960) in California: 10 percent emotionally handicapped and 2 or 3 percent in urgent need of help (pp. 2-3).

The Group I subjects fell within these limits and could be considered to be fairly representative of a normal population. This leaves the problem of the "misclassified" Group II subjects to be dealt with.

Rather than calling the instrument a failure because of its leniency to Group II subjects, a discussion of possible reason for this apparent leniency would seem to be more appropriate.

First, it was decided that a normal sample should have 15% or less of its number classified as behaviorally disturbed if it were to be considered representative of a normal population. This condition was met in terms of Group I. It may have been more profitable to talk of Group II in terms of deviating from normalcy, rather than in terms of a prescribed percentage of misclassification.

Greater than 15% of Group II were below Walker's (1970) critical score on all Scales of the Checklist. However, with the exception of Scale II (at 14.89% above the critical score), all the Scales showed at least 30% of the Group above the critical score. This represented between 2 and 3 times the percentage one would expect to be above the critical score if the Group represented a normal population.

How, then, can the large percentage of "misclassified" Group II subjects be explained? Part of the role of the Adaptation program must be to remediate and control problem

behavior. If this were not the case, why would public money be spent in the maintenance of these classes, when other, less expensive means could be used to simply isolate these children? The classes are small, (if classes sampled here are representative; no more than 11 or 12 students are found in each class, with some classes as small as 6 or 7). This allows for considerable amounts of individualized attention and a resulting level of rapport between pupil and teacher that would not be possible in larger, regular classes. Being small, classroom routines and expectations can be more individualized and, thus, more appropriate to the individual needs of the child. Behavior management techniques are possible at a useful level. Frustration can be lowered by adjusting educational demands to the pupil's abilities.

It must be remembered, too, that most of these students have been with their Adaptation teacher, or in an Adaptation class for a period of at least 7 months, (one exception being a child in Adaptation only 2 months) at the time of data collection,up to a period of 2 years, 7 months. This being the case, if a good percentage of Adaptation students were not moving to within normal limits of behavior, as a result of the remedial aspects of the program, questions about the usefulness of the Adaptation program would ensue. It is, after all, an implicit aim of Special Education

program to re-integrate the exceptional child to the regular stream at the earliest possible opportunity.

Within the limits of the original hypotheses, the instrument has failed to accurately discriminate between Groups I and II because greater than 15% of Group II subjects were rated below the various critical scores on the Checklist. Upon closer investigation it would appear that the limits of the hypotheses were too restrictive. On all but Scale II, more than twice the number expected to be rated as disturbed, in a normal population, were rated as behaviorally disturbed. That this number was not greater may be more in the way of a positive reflection of the Adaptation program, than a negative reflection on the usefulness of the Checklist.

When the mean scores of 47 regular class boys were compared with the mean scores of 47 Adaptation class boys, it was found that these scores generally showed that the groups were significantly different. It was also determined that these two sub-groups were similar to the groups from which they were drawn. Two exceptions to the general differences were noted.

As before, Scale II was the weakest indicator of differences between these groups with no statistically significant differences being noted. Scale V (Immaturity) showed differences significant beyond the .05 level of confidence but not beyond the .01 level. Very little can be said about Scale II differences except that here either the

two groups were similar or the teacher variable accounted for the low apparent differences. Scale V differences, or only moderate presence of differences, may be explainable in terms of a widely held notion that girls tend to mature earlier than boys. The large number of girls in Group I, compared to Group II, may have added to the apparent difference between the groups when full groups were compared, and the absence of girls in this particular analysis could have accounted for the lower order differences here on this scale.

The matched group comparison results tend to be supportive of hypotheses #5 and #6.

As revealed in the hierarchal analysis following Winer's (1962) approach, the Checklist was not immune from individual teacher differences as was evidenced by the moderate but significant teacher effect on both the Full Scale and on Scale II. In the case of the Full Scale teacher effect, the problem was not acute because of the size of the main effect differences, i.e. the difference between regular class students and Junior Adaptation students, and the differences could be discussed as being real despite the teacher effect. Such cannot be said about Scale II (Withdrawal). The main effect differences, though greater than differences due to teacher effect, must be considered as being seriously compromised by the differences attributable to differences among the teachers. As a result, it is

difficult, if not impossible, to attach interpretation to the scale.

The question must be raised concerning the apparent weakness of Scale II (Withdrawal). As noted in the results, not only was its value questionable as the result of the hierarchal analysis, but it appeared as the most inconclusive scale on all of the analyses applied to it and the other scales. Louttit (1957) shed some light on a probable cause for the weakness of this scale:

> Perhaps one of the least disturbing patterns of behavior . . . is that of shyness, seclusiveness or withdrawal. Such behavior is generally not regarded as a serious behavior problem, as indicated by parent or teacher judgments of severity of behavior problems . . . . It is of interest to note that in the cases studied by Martens and Russ (1932), 42 percent of the problem children and 52 percent of the non-problem children were found to be shy and bashful. This particular contrast further suggests that children who meet situations by withdrawal are not likely to be thought of as problems (p. 272).

The percentage of subjects in the present study rated as seriously withdrawn (5.32% of Group I and 14.89% of Group II) came no where near that found by Martens and Russ (1932). Even the number and percentage of subjects rated as exhibiting any form of withdrawal was relatively low (20 subjects or 21.27% of Group I and 34 subjects or 36.17% of Group).

Based on Louttit's (1957) comments, the findings of Martens and Russ (1932) and the results of this study, it appeared that the weakness of the scale rating withdrawal

stemmed from the fact that withdrawal is subject to varying degrees of perceived severity by teachers and is generally not considered to be a particularly serious problem by raters. This notion was further supported by the fact that of the 50 items on the WPBIC only 5 measured this particular scale whereas between 10 and 14 items per scale were represented on the remaining scales. This small number of items could also be considered as a source of weakness of this scale. Thorndike (1971) and Ferguson (1971) suggest that to increase the reliability of a test, one can lengthen it. Since this appeared to be a problem in regard to Scale II, it is conceivable that by increasing the number of items on this scale, some increase in its reliability, seen in a reduction of teacher related variance, could be expected. This, however, would be likely to have little appreciable effect on the number of subjects identified as severely withdrawn, if the findings of Louttit (1957) and Martens and Russ (1932) are true.

It appears, then, that Scale II, because of its susceptability to teacher differences and the fact that withdrawal is often not seen as a problem, was a primary source of weakness in the WPBIC.

Aside from the effect of teacher variance on Scale II, variance other than that due to real differences between groups did not appear to be a major factor. No significant differences in variance were shown to be attributable to

inter-school differences or differences relating to variations of treatment across the schools. In all scales except Scale II, the treatment variance (differences between Group I and Group II) were overwhelmingly the major source of variance. As a result, it is felt that the comments made about the value of the WPBIC based on the other analyses performed on the instrument are reasonably valid.

Walker's (1970) estimate of the reliability of the instrument, at .98, appeared very impressive, especially when compared to the estimate of .8598 for the present study. However the difference in reliability may have been a function of group size (534 v.s. 188) rather than a function of differences between raters and subjects across the two studies. As Ferguson (1971, p. 373) suggested, by increasing the group size, one can expect an increase in the estimate of the reliability of an instrument. Because of this fact an estimate of reliability of .8598 for 188 subjects may be considered as adequate. The purpose of the study was to determine if differences between groups existed, so it was, in fact, a group survey. According to Thorndike (1971):

> If a test is designed for individual diagnosis, the test planner may wish to assure a reliability of .90, whereas for group survey purposes he may tolerate a reliability of .75 to .80 (p. 71).

With the present size of the sample, the estimate of reliability of this study fell between that acceptable

for a group survey and that acceptable for individual diagnosis. Based on this, the data from the present appeared acceptable for discussion on the basis of a group survey. It is to be remembered that the WPBIC would not be the only criterion used in identifying students with behavior problems. In this regard, the estimate of reliability suggested that the instrument had some value as a tool in the initial diagnosis of behavior disorders.

The comparison of Group II to Walker's (1970) experimental group of 38 subjects yielded rather interesting results. It appeared, in those results, that Group II and Walker's experimental group could have come from the same population because of the close similarity of the mean scores and variances. Apparently, the two groups reflected the same overall behavior tendencies, a fact that could be used to support the generalizability of the instrument across populations. The probability of these similarities occurring due to chance would be very remote so it must be assumed that the subjects in the two groups were displaying generally the same types of behaviors.

The two groups matched for age and sex were compared to their source groups to determine if, in fact, they could be said to represent the source groups. The results indicated similarity between matched and source groups so statements regarding these groups could be considered to apply to the source groups. As noted above, the matching

indicated a lessening of differences between Groups I and II in terms of immaturity, and, as also noted above, this could have been due to the absence of female subjects who may have shown more relative maturity than their counterparts in Group I especially.

The results of the correlation between age, time in Adaptation class and checklist scores were somewhat disturbing in view of what was said earlier regarding the role and purpose of Adaptation classes. The slight negative correlation between age and checklist scores was disappointing due to the expectation that as the subjects grew older they would tend to show less disordered behavior. The coefficient of correlation found between these factors was too small at -.143 to allow for definitive statements. The fact that it was a negative correlation was promising because that trend, at least supported the expectation of the direction of the correlation. However, it must be assumed that problem behavior is not age restricted to any significant degree.

The really disturbing result in this analysis, was the lack of clear direction or tendency based on the length of time spent in Adaptation classes. If the Adaptation class were fully meeting the needs of behaviorally disturbed children, there should have been a fairly high negative correlation between the length of time spent in Adaptation classes and the checklist scores. However, with a correlation coefficient of -.054, although the direction of the

coefficient was acceptable, its strength was not. The lack of a high negative correlation, here, could create some concern as to the effect of the Adaptation classes on the students involved in them.

Conclusions and Implications

Based on the results and discussion presented here, certain conclusions can be drawn.

First, based on the data, the checklist appeared to be able to detect real differences between groups representing regular and Adaptation classes. The data supported the instrument's efficacy as a group survey instrument. With an increased sample size, one could, though, possibly not legitimately, claim its usefulness as an individual diagnostic tool. The claim made by Walker (1970) of the instrument's usefulness on an individual basis may have been based on a reliability estimate that was enhanced by group size. This is not to say that Walker's claims regarding reliability were false, because even with a considerably smaller group than his, a coefficient of reliability of nearly .86 was obtained. Even with the lesser of the two estimates, that is, the one found in this study, the instrument appeared to show fairly high reliability.

Because of the ability of the instrument to detect differences between groups, and because of the consistently higher mean scores of the Adaptation subjects, it seems

reasonable to conclude that differences existed in the relative nature of these groups. This statement is made with the full awareness that variances among teachers had significant effects on overall variance in the Full Scale and on Scale II. In the case of the Full Scale this effect was deemphasized due to the size of the Treatment (real differences) effect on variance.

Scale II posed serious problems to the Checklist. Based on present data, Scale II (Withdrawal) would seem to have limited, if any, value for the purpose it was assigned. Observed differences were too small and rater (teacher) variations too great to allow this scale to be given any degree of real value. Teachers and other raters should be cautioned that because of these factors, and the low number of items measuring this scale, extreme caution would be advisable in interpreting the results of this scale. This is not to say that the value of the scale is non-existent, but rather, that the results were too inconclusive to merit its consideration as an effective way of determining the presence of withdrawn behavior.

On the remaining scales, the available data suggests that the instrument was valid for the purpose for which it was designed. On each of Scales I, III, IV and V significant differences of a high order were found between the Groups sampled indicating that the checklist was sensitive to differences along the factors as described.

The Adaptation group's similarity to Walker's experimental group could be construed to lend support to the generalizability of the checklist from Walker's group to the present group. The remarkable similarity of means and variances here are unlikely to have been due to pure chance. The similarity would seem to enhance the image of the checklist as a device for detecting behavior problems at the specified age and grade levels.

The results of the analyses comparing age and time in Adaptation classes to checklist scores were inconclusive. No firm or definitive conclusions regarding these analyses are possible. Further research using all adaptation classes locally might provide more meaningful results.

As a whole, then, the WPBIC appeared valid as a group survey instrument to determine the presence or absence of disturbed behavior in individuals within the classroom. As such, it could be useful as part of a screening process to help identify behaviorally disturbed pupils. Because of teacher related variance effects and weaknesses present due to Scale II, its interpretation should be done in general terms, with further study being conducted of each child identified as behaviorally disturbed on the checklist.

Teacher biasing effects must be kept in mind. It is conceivable that a regular class teacher would seek normalcy in her students and an Adaptation teacher would seek behavior problems to be identified. This possibility must be

recognized as a limitation of the study, because it could conceivably have caused at least some of the differences between groups in the study.

It must also be recognized that a full scale score of less than 21 does not mean the absence of problem behavior. On any given scale, a score above the cut-off point indicates a possible problem in that area. As such the instrument can be useful on a scale by scale basis.

The relationship between academic achievement and full scale scores (revised by Spivak and Swift (1973)) were not dealt with due to the diversity of teacher recorded achievement data.

Despite the directions for research still open regarding the WPBIC and its limitations, it appears that the checklist does have a degree of usefulness, as an initial screening device, that could be of benefit in identifying or helping to identify behaviorally disturbed children at the Grades 4, 5 and 6 level in the local school area.

REFERENCES

Allee, J. G. Webster's Dictionary. Farmingham, Mass: the Literary Press, 1958.

Baldwin, A. L. Theories of Child Development. New York: John Wiley and Sons, Inc., 1967.

Bandura, A. Aggression: A Social Learning Analysis. Englewood Cliffs, N. J.: Prentice Hall, Inc., 1973.

Bandura, A., and Walters, R. H. Social Learning and Personality Development. New York: Holt Rinehart and Winston, 1963.

Berkowitz, L. "On not being able to aggress." British Journal of Social and Clinical Psychology, 1966, 5(2), 130-139.

Bowen, E. M. Early Identification of Emotionally Handicapped Children. Springfield, Ill.: Charles C. Thomas, 1960.

Bryan, T. S., and McGrady, H. J. "Use of a teacher rating scale." Journal of Learning Disabilities, 1972 (Apr.), 5(4), 199-206.

Bryan, T. S., and Wheeler, R. "Perceptions of learning disabled children: the eye of the observer." Journal of Learning Disabilities, 1972, 5(8), 37-41.

Bullock, L. M., and Brown, R. K. "Behavioral dimensions of emotionally disturbed children." Exceptional Children, 1972 (May), 38(9), 740-741.

Cowgill, M. L., Friedland, S., and Shapiro, R. "Predicting learning disabilities from kindergarten reports." Journal of Learning Disabilities, 1973 (Nov.), 6(9), 50-54.

Ebbeson, J. A. "Kindergarten teacher rankings as predictors of academic achievement in the primary grades." Journal of Educational Measurement, 1968, 5(3), 259-262.

Eron, L. D., Huesmann, L. R., Lefkowitz, M. M. and Walder, L. O. "Does television violence cause aggression?" American Psychologist, 1972, 27, 253-263.

Ferguson, G. A. Statistical Analysis in Psychology and Education (3rd Ed.). Toronto: McGraw-Hill, Inc., 1971.

Freehill, M. F. Disturbed and Troubled Children. Flushing, N. J.: Spectrum Publications, Inc., 1973.

Garner, J. and Bing, M. "The elusiveness of Pygmalion and differences in teacher pupil contacts." Interchange, 1973, 4(1), 34-42.

Garrett, H. E. Statistics in Psychology and Education. New York: David McKay Co., Inc., 1962.

Grgin, T. "Rigor-Leniency in male and female examinations in initial grades of school." Psihologija, 1969, 2(1), 303-307.

Hammett, B. C. and Batchelor, H. C. "Problems in measuring children's disturbed behavior." in Frazier, C. A. (Ed.), Is it Moral to Modify Man? Springfield, Ill.: Charles C. Thomas, 1973.

Hartlage, L. C. and Lucas, D. G. "Group screening for reading disability in first grade children." Journal of Learning Disabilities, 1973 (May), 6(5), 317-321.

Kelley, T. L. Interpretation of Educational Measurements. Yonkers, N.Y.: World Book Co., 1927.

Lambert, N. M. and Hartsough, C. S. "Scaling behavioral attributes of children using multiple teacher judgments of pupil characteristics." Educational and Psychological Measurement, 1973, 33(4), 859-874.

Lefkowitz, M. M., Walder, L. O., Eron, L. D. and Huesmann, L. R. "Preference for televised contact sports as related to sex differences in aggression." Developmental Psychology, 1973, 9(3), 417-420.

Lessler, K. and Bridges, J. S. "The prediction of learning problems in a rural setting: Can we improve on readiness tests?" Journal of Learning Disabilities, 1973 (Feb.), 6(2), 90-94.

Lindquist, E. F. Educational Measurement. Washington, D.C.: George Banta Publishing Co., 1950.

Louttit, C. M. Clinical Psychology of Exceptional Children (3rd Ed.), New York: Harper and Row, Publishers, 1957.

Maes, W. R. "The identification of emotionally disturbed elementary school children." Exceptional Children, 1966, 32(9), 607-609.

Maguire, M. S. "A self inventory scale of adolescent symptomology based on the Devereux Adolescent Behavior Scale." Adolescence, 1973 (Sum.), 8(30), 277-284.

Martens, E. H. and Rus, H. "Adjustment of behavior problems of school children. A description and evaluation of the clinical program in Berkeley, Calif." U.S. Official Education Bulletin, 1932, 18.

Piliavin, I. and Briar, S. "Police encounters with juveniles." American Journal of Sociology, 1964, 70, 206-211.

Richards, H. C. "Peabody Picture Vocabulary Test and teacher rated verbal behavior of slum preschoolers." Psychological Reports, 1973, 32(1), 185-185.

Richmond, B. O. and Dalton, J. L. "Teacher ratings and self concept reports of retarded pupils." Exceptional Children, 1973, 40(3), 173-183.

Sibley, S. A., Abbott, M. S. and Cooper, B. P. "Modification of the classroom behavior of a disadvantaged kindergarten boy by social reinforcement and isolation." Journal of Experimental Child Psychology, 1969, 7, 203-219.

Spivak, G. and Swift, M. S. "The classroom behavior of children: A critical review." Journal of Special Education, 1973, 7(1), 55-89.

Swift, M. S. and Spivak, G. "Clarifying the relationship between academic achievement and overt classroom behavior." Exceptional Children, 1969, 36(2), 99-104.

Thorndike, R. L. Educational Measurement (2nd. Ed.). Washington, D.C.: American Council on Education, 1971.

Walker, H. M. Walker Problem Behavior Identification Checklist (Manual), Los Angeles: Western Psychological Services, 1971.

Wang, M. C. "The accuracy of teacher's predictions on children's learning performance." Journal of Educational Research, 1973 (July-Aug.), 66(10), 462-465.

Werry, J. S. and Quay, H. C. "Observing the classroom behavior of elementary school children." Exceptional Children, 1969 (Feb.), 35(6), 461-470.

Wilborn, B. L. "The relationship between teacher attitudes and teacher ratings of pupil behavior." Dissertation Abstracts International, 1972 (Mar.), 32(9a), 4974.

White, M. A. and Harris, M. H. "Mental illness in relation to pupil population." in The School Psychologist. New York: Harper and Row, Publishers, 1961.

Winer, B. J. Statistical Principles in Experimental Design. New York: McGraw-Hill Book Company, Inc., 1962.

Appendix 1

Manual of the Walker Problem Behavior Identification Checklist